Modelos de controlo e alarmística na gestão de ativos fixos

Raquel Pinheiro Leiras,
Engenharia Matemática,
Departamento de Matemática,
2017

Orientador Científico
Prof.ª Doutora Rita Gaio, Docente, Faculdade de Ciências da Universidade do Porto

Orientador de Estágio
Dra. Sílvia Maia, Coordenadora- Desenvolvimento, Controlo e Report, SONAE

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

À Prof.ª Doutora Rita Gaio, pela disponibilidade para orientar este trabalho, pelo acompanhamento, paciência e dedicação que sempre demonstrou para comigo.

À equipa da Gestão de Ativos Fixos da SONAE pelo acolhimento, integração e disponibilidade que apresentaram desde o primeiro momento. Em especial à equipa do Desenvolvimento, Controlo e Reporte por me acolherem e me apoiarem em todas as dificuldades ocorridas no projeto.

À Sílvia Maia e à Susana Silva por estarem sempre presentes, por me cederem toda a informação, pelo apoio, pelo companheirismo e acima de todo por atravessar, em conjunto, todas as dificuldades encontradas ao longo do estágio.

Ao grupo de todos os almoços e lanches, pelo carinho, pelos sorrisos, por todos os abraços, pelo apoio demonstrado e por me acolherem da melhor forma.

Aos meus pais e irmãos, sempre presentes em todas as etapas da minha vida, que com paciência me dão ânimo para ultrapassar todas as adversidades e apoio incondicional. Também um agradecimento à restante família, padrinhos, tios, primos, por estarem sempre presentes, manifestando sempre boa disposição.

Aos meus amigos que o Porto me ofereceu, em especial ao Camacho, ao Ricardo, à Maria e ao Tiago, com quem partilhei momentos e experiências incríveis, e também as dificuldades no desenvolvimento do relatório.

A todos os meus amigos de longa data que representam uma segunda família para mim e que nunca me deixaram desanimar. Em especial, à minha querida Sara que está sempre presente em todos os momentos.

A todos aqueles não mencionados que contribuíram de alguma forma para finalização de mais uma etapa, obrigada.

Resumo

A gestão de ativos fixos é um processo de identificação, contabilização e controlo do equipamento a partir dos ativos de forma coordenada e integrada. Envolve as compras, logística e contabilização de todos os ativos fixos de uma empresa.

O objetivo deste estudo é construir um sistema de controlos e alertas para a deteção de desvios significativos referentes aos volumes dos ativos fixos, por loja, para cada uma das insígnias do grupo SONAE. O estudo centrou-se em ativos fixos tangíveis etiquetáveis, pois são os que têm maior importância a nível de controlo nas lojas por serem alvo de auditorias. Os dados utilizados na presente análise são referentes a ativos fixos tangíveis, que permanecem no património por um longo período de tempo (superior a um ano). Os dados são referentes à insígnia *Y*.

A nível metodológico, foram aplicadas ferramentas de *data mining*, mais concretamente o clustering e a deteção de outliers.

O método de clustering tem como objetivo dividir as observações (as lojas) em grupos homogéneos, denominados clusters. A análise de clusters é um procedimento da estatística multivariada em que a classificação depende da medida de semelhança e do método usado para agrupar os dados. Os métodos clustering podem ser hierárquicos ou não hierárquicos.

Para a identificação de lojas atípicas recorreu-se a diferentes metodologias de deteção de outliers. Usualmente, denomina-se por outliers as observações que diferem substancialmente dos restantes elementos da amostra considerada. Neste trabalho, foram considerados métodos estatísticos, métodos baseados na proximidade e métodos baseados em clustering.

Com o objetivo de resolver este projeto foram abordados três modelos, cada um devolvendo uma lista de lojas outliers. De seguida, foi efetuada uma análise comparativa das diferentes abordagens a fim de identificar qual a mais adequada para o problema em questão. Das diferentes abordagens aplicadas à base de dados, a que proporcionou melhores resultados foi a abordagem mais completa ao nível das técnicas aplicadas. Nesta abordagem aplicou-se o clustering, com o intuito de agrupar as lojas conforme o número de equipamentos e o seu valor; posteriormente aplicaram-se a cada grupo técnicas de deteção de outliers. Por fim, foram

analisadas as lojas outliers. Estas lojas têm uma maior importância para o departamento de "Gestão de Ativos Fixos", uma vez que são lojas que têm demasiados equipamentos para a sua dimensão.

Palavras-Chave: Ativos Fixos; Gestão de Ativos; Data Mining; Clustering; Outliers.

Abstract

The management of fixed assets is a process of identification, accounting, and control of the equipment from the enterprise assets in a coordinated and integrated way. It intends to involve the purchases, logistics and accountings of all the fixed assets of the company.

The objective of this study is to create a system of control and alert for the detection of significant deviations from the volumes of fixed assets, per store. The study focused on labelled tangible assets, since they are the ones of most importance on a store mostly due to the existence of periodic audits.

The data used in the present analysis refers to the tangible fixed assets, that usually remain in store for a long period of time (over one year). The data refer to the Y insignia.

For this analysis, we applied *data mining* tools, more specifically clustering as also detection of outliers.

Clustering has the objective of dividing the set of observations (the stores) into homogeneous groups, called it clusters. Cluster analysis is a procedure of *multivariate statistics*, whose classification depends only on the measure of similarity used to group the data. Clustering methods are generally classified into hierarchical and non-hierarchical.

To finalize, the detection of outliers intended to return the stores whose number of equipments are atypical by comparison with others. Outliers are data that seem to deviate significantly from the remaining data of the sample. From a statistical point of view, we have used methods, methods based on proximity and methods based on clustering. Three methods were applied. Subsequently, a comparative analysis of the different approaches was carried out in order to identify which one is most appropriate for the problem in question. From the different approaches applied to the database, the one that provided better results corresponded to the most complete methodology. In this approach, a clustering technique was applied, with the objective of grouping the stores accordingly to the number of equipment and its value; afterwards, techiniques for outliers detection were applied within each group. The outlier stores were finally analyzed; these stores have a large importance for the department of "Fixed Assets Management"since they have too many equipments for their size.

Keywords: Fixed Assets; Asset Management; Data Mining; Clustering; Outliers

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A maioria das empresas, sejam elas de maior ou menor dimensão, detêm no seu ativo uma parte correspondente a **A**tivos **F**ixos **T**angíveis (AFT). Os ativos fixos tangíveis são recursos que uma empresa detém, com caráter de permanência ou continuidade, não se destinando a ser vendidos ou transformados no decurso das suas atividades normais. São ativos destinados a serem utilizados na produção ou fornecimento de bens e serviços ou para fins administrativos. Estes ativos estão registados na conta 43 – ativos fixos tangíveis, do **S**istema de **N**ormalização **C**ontabilística (SNC) [2]. O SNC veio revogar o POC, **P**lano **O**ficial de **C**ontabilidade, e consiste num conjunto de normas de contabilidade, relato financeiro e de normas interpretativas.

Por muito pequena que seja a quantidade detida de AFT, ela irá ter sempre uma implicação direta nos resultados da organização. Como tal, as decisões de investimento e a gestão económica de AFT são processos muito importantes dentro de uma organização, dado os impactos que têm na rentabilidade presente e futura da empresa. Por essa razão, o processo de gestão de AFT deve ser tão valorizado como qualquer outro.

Este trabalho surge na sequência de um estágio curricular do Mestrado em Engenharia Matemática que decorreu na empresa SONAE, no departamento de "Gestão de Ativos Fixos", com a duração de seis meses. Esse departamento é constituído por quatro equipas que estão interligadas.

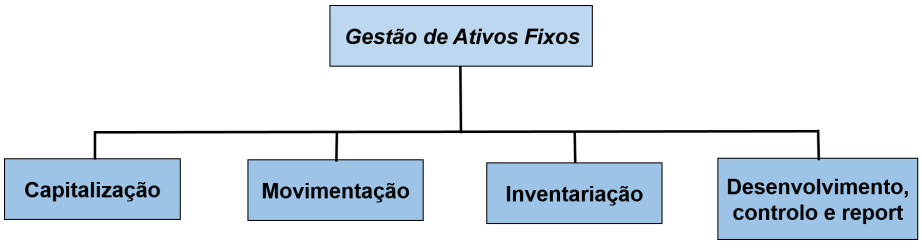

Figura 1.1: Representação do enquadramento das equipas.

Na equipa da **Capitalização** a missão é garantir de forma eficiente a capitalização dos ativos fixos disponibilizando informação rigorosa e atempada, nomeadamente quando o equipamento está instalado ou pronto a ser colocado em serviço, o ativo deve ser capitalizado no sistema *SAP* - consiste num Sistema Integrado de Gestão Empresarial - depois de se reunir toda a faturação referente ao ativo.

Caso tenha que ser dada a baixa de um ativo ou, então, caso esse tenha que ser vendido ou transferido, o responsável pelo local, ou seja o gerente de loja, onde o ativo se encontra deve reportar essa informação à equipa da **Movimentação**.

A equipa da **Inventariação** é a equipa que faz a auditoria e o inventário às lojas, isto é verifica se os ativos estão lá e no caso de haver novos equipamentos coloca a etiqueta correspondente a cada ativo. A equipa de inventariação terá de ir à loja num curto período de tempo quando a loja sofre uma abertura, encerramento, remodelação, transferência ou uma pequena intervenção.

Por último, a equipa **Desenvolvimento, controlo e report** tem a missão de promover a eficiência administrativa e o controlo interno relacionados com a gestão de ativos fixos, assegurando os desenvolvimentos necessários e o suporte aplicacional e garantindo acompanhamento e o envolvimento de todas as equipas da "Gestão de Ativos Fixos".

O objetivo do estudo consiste em implementar modelos de controlo e de alertas para a deteção de desvios significativos do volume de ativos fixos. Pretendem-se identificar as lojas cujo número de equipamentos, por metro quadrado, é atípico.

Os modelos focaram-se nos equipamentos etiquetáveis contidos nos equipamentos básicos, isto é máquinas, ferramentas, equipamentos de decoração e outros bens com os quais se realiza a extração, transformação e elaboração dos produtos ou a prestação dos serviços, pois são equipamentos de foco numa auditoria, ou seja, são equipamentos de controlo.

A elaboração dos modelos baseou-se em três abordagens, todas integradas na área de *data mining*.

Os dois objetivos das técnicas de *data mining* na prática são a **previsão** e a **descrição**. A previsão envolve o uso de algumas variáveis ou campos da base de dados para prever valores desconhecidos ou futuros de variáveis de interesse. A descrição consiste em encontrar padrões que descrevem os dados.

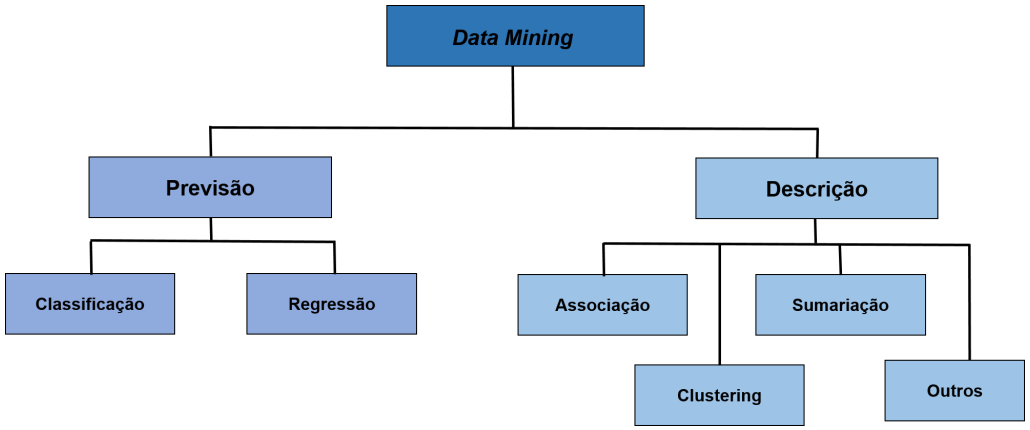

Figura 1.2: Tarefas de *data mining*. [27]

Na literatura [14], as tarefas que aparecem associadas ao *data mining* são: classificação, regressão, sumariação, associação ou dependência, clustering e deteção de outliers.

As ferramentas de *data mining* aplicadas no trabalho presente foram o clustering e a deteção de outliers.

O clustering, ou também segmentação, é um método de segmentação de dados que partilham tendências e padrões semelhantes. O clustering tem como objetivo dividir as observações em grupos - os clusters. De um modo geral no clustering as várias observações que possuírem uma maior semelhança ficam no mesmo cluster e aquelas que apresentam maiores disparidades ficam em clusters distintos.

A análise classificatória de clustering divide-se, genericamente, em dois conjuntos de métodos: os métodos hierárquicos e os métodos não hierárquicos.

De seguida, aplicaram-se técnicas de deteção de outliers, com o intuito de retornar as lojas cujo número e valor dos equipamentos fossem muito elevados ou muito reduzidos, comparado com as demais. As técnicas aplicadas à base de dadas são técnicas não supervisionadas, que se dividem em: técnicas baseadas em métodos estatísticos, técnicas baseadas em clustering e técnicas baseadas na proximidade.

Por fim, foram analisadas as lojas outliers retornadas pelas três abordagens. Estas lojas têm uma maior importância para o departamento de "Gestão de Ativos Fixos". As equipas analisam as lojas outliers de forma a perceberem quais as razões destas conterem muitos/poucos equipamentos.

1.1 Estrutura da dissertação

A presente dissertação encontra-se subdividida em 5 capítulos. No primeiro, descreve-se o problema em estudo e as metodologias estatísticas usadas na sua resolução. No segundo, encontram-se explanados os diversos métodos de clustering, bem como as medidas de semelhança e a avaliação dos clusters. No capítulo 3 é apresentado o tema da deteção de outliers, iniciando-se com a definição e identificação dos diversos tipos de outliers, bem como as principais abordagens existentes nesta área. O resultado da aplicação dos métodos selecionados sobre o conjunto de dados criado é apresentado no capítulo 4. Por fim, são apresentadas as conclusões obtidas ao trabalho desenvolvido e ainda no capítulo 5, são sugeridas outras possíveis abordagens.

Capítulo 2

Clustering

Das tarefas de *data mining* destaca-se o **clustering**, que permite encontrar subconjuntos de dados semelhantes entre si. Esta tarefa vai de encontro ao problema abordado na tese de agrupar lojas em cada insígnia de acordo com características comuns quanto aos seus ativos fixos.

Os métodos de clustering são úteis para agrupar os dados, onde estes grupos são denominados por **clusters**. Os clusters não são conhecidos previamente. De um modo geral no clustering as várias observações são 'comparadas' e aquelas que possuírem uma maior semelhança ficam no mesmo cluster e aquelas que apresentam maiores disparidades ficam em clusters diferentes, como mostra a figura 2.1Exemplo da aplicação da técnica clustering. figure.caption.11. Um bom método de agrupamento forma grupos cuja semelhança entre elementos no mesmo grupo é elevada e a semelhança entre elementos de grupos diferentes é reduzida.

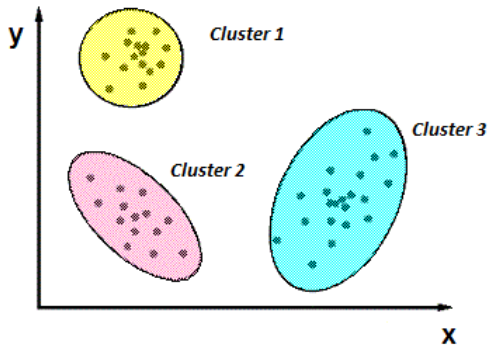

Figura 2.1: Exemplo da aplicação da técnica clustering.

Note-se que ao aplicar o clustering pode-se considerar dois casos extremos: cada observação ser um cluster ou todas as observações formarem apenas um cluster.

A análise classificativa divide-se, genericamente, em dois conjuntos: os **métodos hierárquicos** e os **métodos não hierárquicos**.

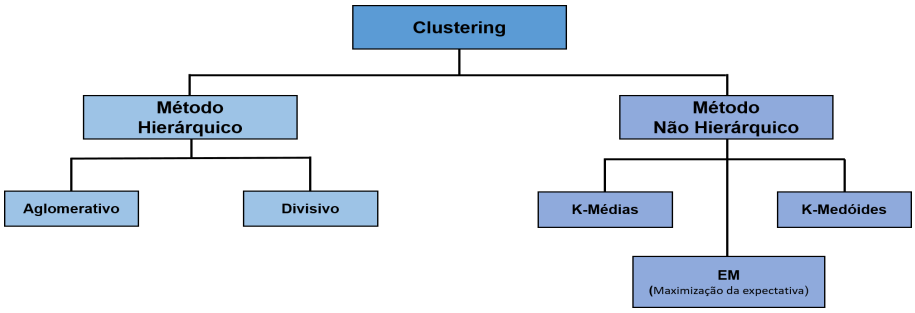

Figura 2.2: Taxonomia de clustering.

O método hierárquico consiste numa série de sucessivas fusões ou sucessivas divisões dos elementos, onde os elementos são agregados ou desagregados, respetivamente.

O algoritmo do método não hierárquico, isto é, clusterização por particionamento é frequente definir à partida o número K de clusters que se pretende criar, onde K é definido pelo utilizador. O objetivo será determinar a classificação das n observações em K classes que otimize algum critério de homogeneidade interna e heterogeneidade externa.

Note-se que os algoritmos de agrupamento não são determinísticos, o que significa que ao usar diferentes condições iniciais obtêm-se resultados muito diferentes. A análise de agrupamento pode, portanto, conduzir a vários conjuntos de clusters. Algumas das limitações deste tipo de análise devem-se ao facto de: não detetar o número de clusters existentes por natureza na amostra; não indicar o melhor particionamento; nem sempre criar grupos facilmente identificáveis e de dimensão razoável; e não ter em consideração as relações existentes entre as variáveis.

Uma definição mais formal de clustering dada por Hruschka et al., em 2013. Considere-se um conjunto de n observações, $X = x_1, x_2, ..., x_n$, onde cada $x_i \in \mathbb{R}^p$ é um vetor de atributos constituídos por p medidas reais. Pretende-se agrupar as observações em K clusters disjuntos, $C = c_1, c_2, ..., c_K$, de forma a respeitar as seguintes condições [10]:

1. $c_1 \cup c_2 \cup ... \cup c_K = X$;

2. $c_i \neq \emptyset$; e

3. $c_i \cap c_j = \emptyset, \quad com\ i \neq j$.

Realça-se que, por essas condições, uma observação não pode pertencer a mais de um cluster (grupos disjuntos) e que cada cluster tem que ter pelo menos uma observação.

2.0.1 Medidas de semelhança

Um conceito muito importante e muito utilizado em *data mining* é a noção de semelhança. Goldschmidt e Passos, em 2015, afirmam que "uma vez que o conjunto de dados pode ser interpretado como um conjunto de pontos num espaço K-dimensional, o conceito de semelhança entre dois pontos pode ser traduzido como a distância entre esses pontos" [15]. A medida de semelhança é crucial para a construção de um cluster, pois, se dois padrões são semelhantes de acordo com algum critério utilizado, então serão agrupados no mesmo cluster, caso contrário, serão agrupados em clusters distintos.

O conceito de distância é formalizado como sendo, E o conjunto de pontos e x, y e z elementos quaisquer de E. A distância entre um par de pontos de E é uma função $dist : E \times E \longrightarrow \mathbb{R}_0^+$, que verifica as seguintes propriedades [15]:

1. $dist(x, y) \geqslant 0$, positividade;

2. $dist(x, x) = 0$;

3. $dist(x, y) = dist(y, x)$, simetria; e

4. $dist(x, y) \leqslant dist(x, z) + dist(y, z)$, desigualdade triangular.

As distâncias mais utilizadas no estudo de uma base de dados com variáveis quantitativas são:

1. Distância Euclidiana
$$dist(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}; \tag{2.1}$$

2. Distância de Manhattan:
$$dist(x, y) = \sum_{i=1}^{n} \mid x_i - y_i \mid; e \tag{2.2}$$

3. Distância de Chebychev:
$$dist(x, y) = max \mid x_i - y_i \mid, \tag{2.3}$$

onde x_i e y_i são as coordenadas dos pontos x e y, respetivamente, e n é a quantidade de coordenadas de cada ponto.

A distância utilizada nos métodos usados para a elaboração dos modelos alarmísticos é a distância ***Euclidiana***, devido a ser uma das distâncias mais comuns e mais simples.

2.0.2 Método hierárquico

Os métodos hierárquicos englobam técnicas que apresentam os dados sob a forma de uma hierarquia. Existem dois métodos para criar uma hierarquia: **hierárquicos aglomerativos** e **hierárquicos divisivos**.

O método hierárquico aglomerativo constrói uma hierarquia de baixo para cima, isto é, considera-se n clusters, onde n é o número de observações, cada par de clusters é interativamente fundido num novo, decrescendo o número de clusters na ordem de um.

O método hierárquico divisivo consiste em construir uma hierarquia de cima para baixo, isto é, partir de um único cluster, englobando todas as n observações, e iniciar um processo de subdivisões sucessivas.

A seleção do par de clusters que deve ser associado (nos aglomerativos) ou do cluster que deve ser separado (nos divisivos) é feita pelo valor da função objetivo obtida pelo agrupamento ou pela divisão.

Os resultados dos métodos hierárquicos são geralmente apresentados em dendogramas, onde representa as relações entre os clusters e a ordem pela qual os clusters foram fundidos (métodos aglomerativos) ou divididos (métodos divisivos).

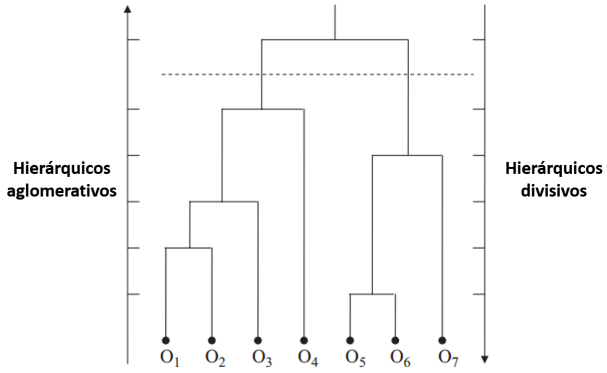

Figura 2.3: Exemplo de um dendograma. A direção de agrupamento do método divisivo é oposta ao do método aglomerativo. Dois clusters são obtidos cortando o dendograma a um nível apropriado. [37]

Método hierárquico aglomerativo

Os métodos hierárquicos aglomerativos são, geralmente, mais utilizados do que os métodos divisivos [29].

A ideia básica do algoritmo do método hierárquico aglomerativo é bastante simples, uma abordagem de baixo para cima. O algoritmo começa com n clusters, cada um deles contendo

uma observação dos dados originais. Para cada par de clusters, a distância (ou semelhança) entre eles é calculada, de acordo com uma determinada medida de distância. Isso produz uma matriz de distâncias $n \times n$, cuja célula (i, j) contém o valor da distância entre os clusters i e j. Em seguida, o algoritmo funde o par mais próximo de clusters, e uma nova matriz de distâncias $(n-1) \times (n-1)$ é formada. Esse processo é executado iterativamente até que haja apenas um cluster, contendo todas as observações do conjunto de dados original.

As principais medidas de proximidade usadas nos métodos hierárquicos aglomerativos, para determinar quais os grupos a serem fundidos, são [26]:

- **Ligação simples** funde grupos baseados na distância mínima entre dois grupos (figura 2.4Diferentes formas de ligação para agrupamento hierárquico. (a)Ligação simples, (b)Ligação completa e (c)Ligação média. [26] figure.caption.15, portanto a distância entre os clusters R e Q é definida por:

$$d_S(R, Q) = min_{i \in R, j \in Q}\ dist(i, j), \tag{2.4}$$

  onde $dist(i, j)$ é a distância entre a i-ésima e a j-ésima observação.

- **Ligação completa** agrupa baseado na distância máxima entre dois grupos, ver figura 2.4Diferentes formas de ligação para agrupamento hierárquico. (a)Ligação simples, (b)Ligação completa e (c)Ligação média. [26] figure.caption.15. Em outras palavras, a distância entre os clusters R e Q é definida por:

$$d_C(R, Q) = max_{i \in R, j \in Q}\ dist(i, j). \tag{2.5}$$

- **Ligação média** agrupa com base na distância média de todas as observações de um grupo a todas as observações num outro, ver figura 2.4Diferentes formas de ligação para agrupamento hierárquico. (a)Ligação simples, (b)Ligação completa e (c)Ligação média. [26] figure.caption.15. A distância da ligação média é definida pela seguinte expressão:

$$d_A(R, Q) = \frac{1}{|R||Q|} \sum_{i \in R, j \in Q} dist(i, j) \tag{2.6}$$

  onde $|R|$ é o número de observações no cluster R.

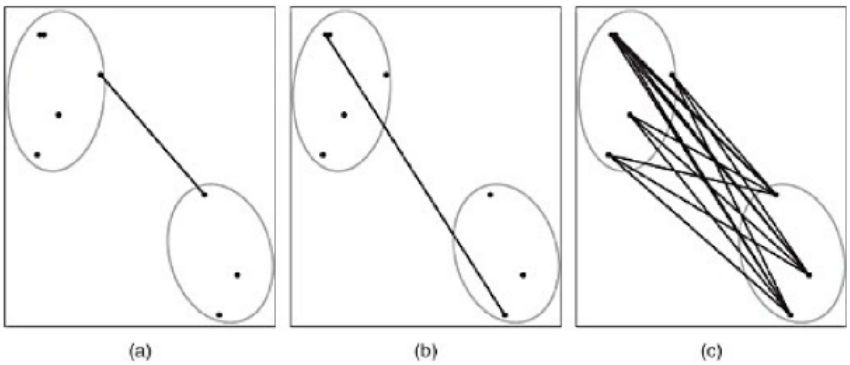

Figura 2.4: Diferentes formas de ligação para agrupamento hierárquico. (a)Ligação simples, (b)Ligação completa e (c)Ligação média. [26]

- **Método de Ward** é dos métodos mais utilizados nos métodos hierárquicos aglomerativos; inicia-se com n clusters, cada um contendo uma observação, e determina-se a soma dos erros quadrados, J_s, para estabelecer os próximos dois grupos a fundir em cada etapa do algoritmo. A soma dos erros quadrados, J_s, é definida (para dados multivariados) por [26]:

$$J_s = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \parallel x_{ij} - \overline{x}_i \parallel^2, \tag{2.7}$$

onde x_{ij} é a j-ésima observação no i-ésimo cluster, $\overline{x}_i$ é a média da amostra para o i-ésimo cluster com n_i observações.

- **Método de centróide:** a distância entre os grupos é a distância entre os seus centróides, que são os valores médios das observações em relação às variáveis. Cada vez que as observações são agrupadas, um novo centróide é calculado. Tanto este método quanto o de Ward exigem a distância euclidiana. O método de centróide é definido pela seguinte expressão:

$$d_c(R, Q) = dist(\overline{x_R}, \overline{x_D}), \tag{2.8}$$

com $\overline{x_R} = \frac{\sum_{i \in R}}{n_R}$ e $\overline{x_D} = \frac{\sum_{i \in D}}{n_D}$.

Input:

D: base de dados que contém n observações.

Output:

Obtém-se um único cluster, formado a partir da matriz de semelhança.

O algoritmo geral para o agrupamento hierárquico aglomerativo é o seguinte:

1. Inicia com n grupos, cada um com uma única observação;

2. Calcula a matriz simétrica de ordem n da distância entre as observações i e j, $dist_{ij}$;

$$dist_{n\times n} = \begin{bmatrix} dist_{11} & dist_{12} & ... & dist_{1n} \\ ... & ... & ... & ... \\ dist_{n1} & dist_{n2} & ... & dist_{nn} \end{bmatrix}; \tag{2.9}$$

3. Funde os grupos com a distância menor, formado assim (RQ), onde R e Q são grupos;

4. Atualiza a matriz de semelhança entre o novo cluster e todos os restantes; e

5. Repete o passo 2, $N-1$ vezes, até se obter um único cluster.

2.0.3 Método não hierárquico

Em contraste com o método hierárquico, que gera clusters numa série de sucessivas fusões ou sucessivas divisões, o método não hierárquico produz uma partição num número fixo de classes. Mais especificamente, dado um conjunto de pontos da base de dados $x_j \in \mathbb{R}^p, j = 1, ..., n$, os algoritmos de clustering particional têm como objetivo organizá-los em K clusters $c_1, ..., c_K$ enquanto maximizam ou minimizam uma função de critério pré-especificada. Uma dessas funções de critério é a soma dos erros quadrados, J_s que é dada pela expressão 2.8Método hierárquico aglomerativoequation.2.8.

A partição que minimiza o critério da soma dos erros quadrados é considerada ótima e é chamada de partição da variância mínima. Este critério é apropriado para os clusters compactos e bem separados. No entanto, o critério pode ser sensível à existência de outliers e, consequentemente, pode dividir incorretamente um grande cluster em pequenos, como refere Duda et al., em 2001, [9]. Os algoritmos que se destacam nos métodos não hierárquicos são os **K-médias** e os **K-medóides**.

Método K-Médias

O método K-médias foi aplicado por James MacQueen, em 1967. Este método é dos mais populares no clustering, é muito usado na área cientifica e na industrial. Este algoritmo é usado para particionar a base de dados, onde K é a quantidade de centróides (pontos centrais dos cluters) que serão criados e ajudarão a encontrar a similaridade dos dados.

Note-se que os clusters são gerados dependendo da semente (centros iniciais do clusters), o algoritmo gera diferentes clusters, e pré-especificando as sementes torna o algoritmo mais rápido e poderá retornar uma boa solução.

O algoritmo K-médias começa com uma partição inicial das observações em K grupos, com as suas sementes geradas aleatoriamente. Iterativamente a partição é modificada de forma a minimizar a distância de cada observação ao centróide a que pertence. Esse passo gera uma nova partição em que a soma das distâncias tende a ser menor comparando com a partição gerada anteriormente. O passo de gerar a partição de forma a reduzir a soma da distância é repetido até que o valor seja próximo do nulo.

Como foi referido anteriormente, o objetivo é minimizar a variância em cada cluster, ou seja, as observações em cada cluster precisam de estar próximas. Para tal aplica-se a soma dos erros quadrados dada pela expressão 2.8Método hierárquico aglomerativoequation.2.8.

Input:

D: base de dados que contém n observações;

K: número de clusters a formar.

Output:

Um conjunto de K clusters, que minimiza a soma das medidas de dissemelhança de cada observação para o centóide mais próximo.

O algoritmo K-médias exige a pré-fixação dos critérios e, usualmente, o procedimento segue os seguintes passos [11]:

1. Seleciona as observações em K grupos iniciais;
2. Recalcula cada observação no grupo cujo centróide esteja mais próximo;
3. O centróide é recalculado para o grupo que recebeu nova observação e para o grupo que perdeu alguma; e
4. Repete o passo 2 até que não restem mais realocações a serem feitas.

Basicamente, define-se a distância *Euclidiana* para determinar o cluster a que pertence cada observação. Compara-se a distância entre um dos clusters, c_i, e um ponto da base de dados, x_K, que pertence ao cluster mais próximo:

$$l_K(x_K) = arg\ mindist(x_K - c_i), \tag{2.10}$$

onde l_K é a classe do ponto x_K. O algoritmo K-médias tenta encontrar um conjunto de centros dos clusters, de forma a que o erro total seja mínimo.

Note-se que o número de iterações necessárias pode variar numa ampla faixa de algumas a milhares dependendo do número de observações, do número de clusters e da base de dados.

O método K-médias pode ser computacionalmente muito intensivo, dependendo do número de variáveis na base de dados.

Método K-Medóides

O método K-medóides não é muito distinto do método K-médias, a diferença está na representação dos centros dos clusters, ou seja, em vez de utilizar o centróide, este método usa o medóide. Um medóide de um determinado agrupamento pode ser definido como a observação do grupo, cuja soma das distâncias a todas as observações do mesmo grupo seja mínima, ou seja, é um ponto mais centralmente localizado no cluster. As observações, não medóides, são atribuídas ao cluster cujo medóide é mais próximo. Como foi referido anteriormente, o objetivo é minimizar a variância em cada cluster, ou seja, as observações em cada cluster precisam de estar próximas. Para tal aplica-se a soma dos erros quadrados dada pela expressão 2.8Método hierárquico aglomerativoequation.2.8. O objetivo em cada iteração é reduzir o valor de J_s.

O algoritmo funciona do seguinte modo: determina *a priori* K clusters para n observações definindo, arbitrariamente, um medóide representativo para cada cluster. Cada observação é atribuída ao medóide mais semelhante, ou seja, é atribuído ao medóide mais próximo. Os medóides são substituídos iterativamente por um dos não medóides desde que a qualidade da partição seja melhorada, isto é, o valor de J_s, soma dos erros quadrados, seja reduzido em comparação com a iteração anterior. Entre os algoritmos da família dos métodos K-medóides destaca-se o **algoritmo de partição em torno dos medóides - PAM** [5].

Input:

D: base de dados que contém n observações;

K: número de clusters a formar.

Output:

Um conjunto de K clusters, que minimiza a soma das medidas de dissemelhança de cada observação ao medóide mais próximo.

O algoritmo de K-medóides segue os seguintes passos [11]:

1. Seleciona K observações em D, como medóides;
2. Atribui cada observação não medóide ao cluster que contém o medóide mais próximo;
3. Seleciona, aleatoriamente, uma observação não medóide O_{random};
4. Calcula o custo total, S, da troca de O_i com O_{random}, onde S= custo_total_atual - custo_total_anterior;

5. Se $S < 0$, então troca O_i com O_{random}, formando um novo conjunto de K-medóide; e

6. Repete 2 até que não ocorra troca de medóides.

À semelhança do K-médias, a construção dos clusters no K-medóides é diferente conforme a semente (medóide inicial do cluster) usada. Uma das primeiras vantagens do K-medóides referenciada por Kaufman e Rousseeuw, em 1990, é determinar os medóides iniciais de forma mais sofisticada. O esforço adicional na fase de inicialização é contrastado por menores esforços computacionais durante a atualização dos medóides [5].

Note-se que o método K-medóides não é tão sensível a outliers como o método K-médias. No entanto, o grande número de comparações entre pares leva a maiores esforços computacionais e fraca escala. Uma desvantagem, à semelhança com o método K-médias, é o facto destes métodos necessitarem do número K de clusters a formar ser indicado antecipadamente pelo utilizador.

2.1 Avaliação dos clusters

Em termos gerais, existem três abordagens para avaliar os resultados de um processo de clustering: critérios externos, critérios internos e critérios relativos.

Critérios externos: consistem em avaliar resultados do algoritmo de cluster baseado numa estrutura pré-especificada, que é imposta na base de dados e reflete a intuição sobre a estrutura do agrupamento das observações.

Critérios internos: os resultados do algoritmo de cluster são avaliados em termos de quantificações que envolvem os vetores do próprio conjunto de dados (por exemplo, a matriz das distâncias).

Critérios relativos: têm como objetivo encontrar o melhor agrupamento que o algoritmo pode obter sob certas suposições e valores para os parâmetros. O caso mais comum para o índice de avaliação relativo é selecionar o melhor algoritmo de clustering a partir do conjunto de resultados obtidos usando diferentes parâmetros.

As duas primeiras abordagens são baseadas em testes estatísticos. Os índices relacionados com essas abordagens têm o objetivo de medir se o resultado confirma a hipótese pré-especificada.

Nesta secção são abordados dois índices de avaliação dos grupos, o coeficiente de correlação cofenética [34] e a largura média da silhueta [31]. Essas medidas, avaliam o resultado do agrupamento por meio de critérios internos.

2.1.1 Coeficiente de Correlação Cofenética

O **C**oeficiente de **C**orrelação **C**ofenética (CCC), apresentado por Farris, em 1969, é uma medida de validação para métodos hierárquicos. O CCC é usado para medir o grau de ajuste entre a matriz de semelhança original (matriz D) e a matriz resultante da simplificação proporcionada pelo método de clustering hierárquico (matriz C) [34].

O CCC é definido como a correlação entre as similaridades originais em $M = \frac{N(N-1)}{2}$ pares, onde N é o número de observações da base de dados. A semelhança cofenética, c_{ij}, entre dois vetores i e j é a distância em que os dois vetores são fundidos no mesmo conjunto. A distância, com base no clustering, é calculada pela expressão 2.1Medidas de semelhançaequation.2.1.

O coeficiente de correlação cofenética é dado pela seguinte expressão:

$$CCC = \left| \frac{\left(\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^{p} c_{ij} - \mu_p \mu_c\right)}{\sqrt{\left[\left(\frac{1}{M}\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(d_{ij}^{p})^2 - \mu_p^2\right]}\sqrt{\left[\left(\frac{1}{M}\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(c_{ij})^2 - \mu_c^2\right]}} \right|, \tag{2.11}$$

onde μ_p e μ_c são as médias dos elementos das matrizes de semelhança originais e as matrizes cofenéticas, respetivamente, enquanto d_{ij}^{p} e c_{ij} é os (i,j)-ésimos elementos das matrizes de semelhança originais e das matrizes cofenéticas, respetivamente. A concordância entre os dados de entrada e o dendrograma está próxima se o valor do índice estiver perto de 1. Um valor alto para CCC é considerado como uma medida de classificação bem-sucedida.

2.1.2 Largura média da silhueta

O gráfico da silhueta é uma técnica que foi proposta por Rousseeuw, em 1986, para avaliar particionamentos. A silhueta de um cluster reflete a qualidade da alocação das observações no cluster [31]. Isso pode ser considerado uma medida composta de homogeneidade e separação de clusters. As parcelas da silhueta fornecem uma representação gráfica de forma a ser possível observar se os clusters são compactos em comparação com outros.

Para construir o gráfico da silhueta é necessário ter uma partição dos dados (obtida por exemplo da aplicação de uma técnica de clustering) e a distância entre as observações. Cada observação x_i é representada por um valor $s(x_i)$ chamado de silhueta, que é baseado na comparação da 'consistência' e na 'separação' de cada cluster.

Seja x_i qualquer observação do conjunto de dados, e denomina-se por A o cluster ao qual foi atribuída. Quando o cluster A estiver concluído, calcula-se pela seguinte expressão a semelhança média $a(x_i)$ da observação x_i em relação às restantes observações em A:

$$a(x_i) = \frac{1}{n_A - 1}\sum_{j \in A, j \neq i} dist(x_i, x_j) \tag{2.12}$$

onde n_A representa o total de observações contidas no cluster A e $dist(x_i, x_j)$ é a medida de semelhança entre as observações x_i e x_j.

Considera-se qualquer conjunto C, em que $C \neq A$, e calcula-se $dist(x_i, C)$, tal que a medida de semelhança média de x_i, para todas as observações de C, é dado por:

$$dist(x_i, C) = \frac{1}{n_C} \sum_{j \in C} dist(x_i, x_j), \tag{2.13}$$

onde n_C representa o total de observações contidas no cluster C. Depois de calcular $dist(x_i, C)$ para todos os clusters C, seleciona-se o menor desses números e denota-se por:

$$b(x_i) = min\{dist(x_i, C), \forall C \neq A\}. \tag{2.14}$$

O cluster B para o qual este mínimo é atingido (isto é, $d(x_i, B) = b(x_i)$) chama-se o vizinho da observação x_i.

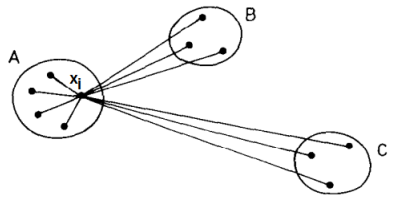

Figura 2.5: Representação dos elementos envolvidos no calculo de $s(x_i)$, onde a observação x_i pertence ao cluster A. [31]

O valor da silhueta da observação x_i é calculado por:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}}, \tag{2.15}$$

em que $-1 \leq s(x_i) \leq 1$ para cada observação x_i e pode ser interpretado da seguinte forma: $s(x_i) \simeq 1$, a observação x_i está bem classificada no cluster A; $s(x_i) \simeq 0$, a observação x_i está entre os clusters A e B; e $s(x_i) \simeq -1$, a observação x_i mal classificada no cluster A e mais próxima do cluster B, quanto mais próximo de 1 melhor é a qualidade da partição.

As silhuetas são mostradas como barras horizontais. Cada observação tem uma silhueta, onde silhuetas largas indicam uma forte semelhança entre as observações dentro do cluster e fraca semelhança entre as observações de outros clusters. Silhuetas negativas sugerem uma má alocação de uma observação no cluster, isto é, um cluster com uma silhueta negativa é mais semelhante a observações de outros clusters do que a observações do seu próprio cluster.

Os valores médios da silhueta podem ser interpretados como se segue na tabela 2.1Valores médios da silhueta. [31] table.caption.19:

$s(x_i)$	**Descrição**
0.71-1.00	Estrutura forte.
0.51-0.70	Estrutura razoável.
0.26-0.50	Estrutura fraca.
$\leqslant 0.25$	Estrutura substancial.

Tabela 2.1: Valores médios da silhueta. [31]

Note-se que se pode usar a silhueta para determinar qual o melhor número de clusters a formar na base de dados.

Capítulo 3

Deteção de *outliers*

Um outlier é um ponto da base de dados que é significativamente diferente dos restantes.

Hawkins, em 1980, definiu outlier do seguinte modo: "Um outlier é uma observação que se desvia das restantes observações, parecendo que foram geradas por um mecanismo diferente" [18].

Ramasmawy et al., em 2000, definem que "Um outlier é uma observação ou um ponto que é consideravelmente diferente ou inconsistente quando comparando com as restantes" [28].

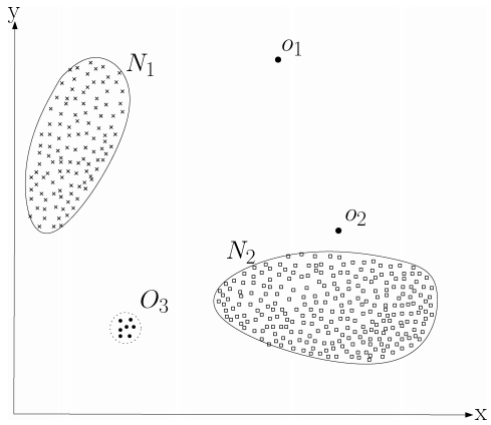

Figura 3.1: Um exemplo simples de outliers num conjunto de dados bidimensionais. [7]

Os outliers podem aparecer isolados das demais observações ou como um conjunto pequeno de pontos, distantes da grande maioria das observações, como se pode observar na figura 3.1Um exemplo simples de outliers num conjunto de dados bidimensionais. [7] figure.caption.20, onde os outliers correspondem aos pontos o_1, o_2 e ao conjunto O_3.

Muitos métodos de *data mining* eliminam os *outliers* como sendo ruídos ou exceções. No entanto, em algumas aplicações, como por exemplo a deteção de fraudes, os valores extremos podem ter mais importância do que os restantes valores.

Na maioria das aplicações, os dados são concebidos por um ou mais processos geradores, que podem refletir a atividade no sistema. Quando o processo de geração se comporta de maneira incomum, surgem os outliers. Portanto, um outlier geralmente contém informações úteis sobre características anormais da atividade no sistema. O reconhecimento de tais características incomuns fornece conhecimentos úteis para aplicações específicas. A título de curiosidade seguem-se os seguintes exemplos de aplicabilidade da deteção de outliers [7]:

- **Sistemas de deteção de intrusão** (IDS): são utilizados em ambientes corporativos, assim como em redes locais, dependendo do modelo aplicado. A funcionalidade principal é reforçar como uma segunda defesa à invasão, atuando posteriormente à constatação da intrusão no sistema.

- **Deteção de fraude:** refere-se à deteção de atividades criminosas que ocorrem em organizações comerciais, tais como bancos, empresas de cartão de crédito, agências de seguros, entre outras. Os usuários mal-intencionados podem ser os clientes reais ou podem-se fazer passar por clientes. A fraude ocorre quando esses usuários consomem os recursos fornecidos pela organização de forma não autorizada.

- **Diagnóstico médico:** em muitas aplicações médicas, os dados são recolhidos de uma variedade de dispositivos, como exames de ressonância magnética, exames de tomografia por emissão de positrões (PET) ou séries de tempo de eletrocardiograma (ECG). Padrões incomuns em tais dados refletem condições de doença.

- **Ciências da Terra:** uma quantidade significativa de dados espaciotemporais sobre padrões climáticos, mudanças climáticas ou padrões de cobertura de terra é retirada através de uma variedade de mecanismos, como satélites ou sensoriamente remoto. As anomalias nesses dados fornecem perceções significativas sobre tendências humanas ou tendências ambientais, que podem ter causado tais anomalias.

Em todas estas aplicações, os dados têm um modelo 'normal' e os outliers são reconhecidos como desvios deste modelo.

3.1 Tipos de outliers

Um aspeto importante das técnicas de deteção de outliers é a natureza da anomalia. Chandola et al., em 2009, classificam os outliers em três categorias [7]: outliers pontuais, outliers contextuais e outliers coletivos.

Outliers pontuais: se uma observação dos dados é considerada anómala em relação às restantes, então a observação é denominada por ponto distinto. Este é o tipo mais simples de outliers e é o foco da maioria das pesquisas sobre a deteção de outliers. Por exemplo, na figura 3.1Um exemplo simples de outliers num conjunto de dados bidimensionais. [7] figure.caption.20, os pontos o_1 e o_2, assim como os pontos na região O_3, estão fora do limite das regiões 'normais' e, portanto, são pontos distintos, uma vez que são diferentes dos restantes pontos. Um exemplo real é a deteção de fraudes dos cartões de crédito.

Outliers contextuais: se uma observação dos dados é aberrante num contexto específico, mas não de outra forma, então é denominada como sendo um outlier contextual. A noção de contexto é induzida pela estrutura da base de dados e deve ser especificada como parte da formulação do problema. Uma observação pode ser um outlier contextual num dado contexto, mas uma observação idêntica pode ser considerada 'normal' num contexto diferente. A figura 3.2Representação de um outlier contextual em t_2 numa série temporal de temperaturas. [7] figure.caption.21 representa um exemplo para uma série temporal de temperaturas que mostra a temperatura mensal numa área ao longo dos últimos anos. Uma temperatura de 5^oC pode ser normal durante o inverno (no tempo t_1), mas o mesmo valor durante o mês de junho (no tempo t_2) seria um outlier.

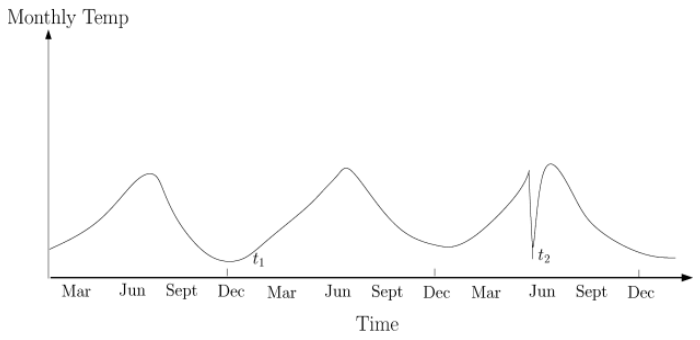

Figura 3.2: Representação de um outlier contextual em t_2 numa série temporal de temperaturas. [7]

Outliers coletivos: um subconjunto de dados forma um outlier coletivo se o subconjunto se desviar dos restantes dados. A figura 3.3Representação de um outlier coletivo correspondente a uma contração prematura atrial numa saída de eletrocardiograma humano. [7] figure.caption.22 é um exemplo que mostra uma saída de eletrocardiograma humano. A região realçada indica um valor atípico porque o mesmo valor baixo existe para um tempo, anormalmente, longo.

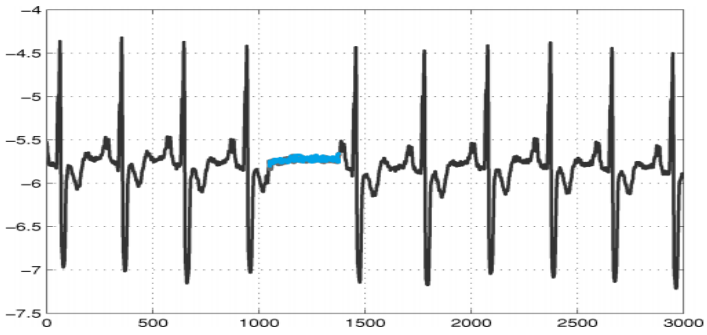

Figura 3.3: Representação de um outlier coletivo correspondente a uma contração prematura atrial numa saída de eletrocardiograma humano. [7]

As técnicas usadas para detetar outliers coletivos são muito diferentes das técnicas de deteção de outliers pontuais e contextuais.

3.2 Técnicas de deteção de outliers

Um fator importante para a deteção de outliers é a presença ou não da classificação do conjunto de dados. Uma classificação é uma informação associada a cada observação da base de dados, indicando se é inlier (normal) ou se é outlier (anómala).

De acordo com a presença ou não desta classificação no conjunto de dados, as técnicas de deteção de outliers podem ser classificadas em três grupos [7]: deteção supervisionada, deteção semi-supervisionada e deteção não supervisionada.

A **deteção supervisionada de outliers** consiste em um conjunto de técnicas que atuam num modo supervisionado, assumindo que as observações da base de dados estão classificadas de 'normais' ou de outliers. Este tipo de dados são úteis em modelos de previsão.

A **deteção semi-supervisada de outliers** consiste em técnicas que atuam num modo semi-supervisionado, assumindo que apenas as amostras 'normais' estão classificadas. Uma vez que não requerem rótulos para a classe de outlier, estas técnicas são mais aplicáveis do que as técnicas supervisionadas. A abordagem utilizada constrói um modelo para a classe correspondente ao comportamento normal e utiliza esse modelo para identificar outliers no conjunto da amostra. Estas técnicas não são muito usadas, principalmente porque é difícil obter um conjunto de dados de treino que cubra todos os possíveis outliers da base de dados.

A **deteção não supervisionada de outliers** consiste em um conjunto de técnicas que atuam num modo não supervisionada, isto é, as amostras não estão classificadas. As técnicas, nesta categoria, fazem a suposição implícita de que os conjuntos de dados 'normais' são mais frequentes do que os outliers. Muitas técnicas semi-supervisionadas podem ser adaptadas para

técnicas não supervisionadas, sendo que a adaptação pressupõe que os dados de teste contenham poucos outliers.

3.2.1 Saída de deteção de outliers

Um aspeto importante para qualquer técnica de deteção de valores atípicos é a maneira pela qual os outliers são descritos. Tipicamente, as saídas produzidas pelas técnicas de deteção de outliers são de um dos seguintes tipos [7]:

- ***Scores*:** atribuem um score de outliers a cada observação e uma observação é ou não considerada um outlier dependendo do grau de anormalidade. Constrói-se uma lista com os graus de outliers de cada observação dispostos por ordem decrescente; serão consideradas possíveis observações outliers aquelas com maiores valores no score. Para uma analise posterior pode-se optar por analisar os poucos valores excecionais ou usar um limite para o grau de anomalia para selecionar possíveis outliers.
- **Classificação:** técnicas nesta categoria atribuem uma classe ('normal' ou outlier) a cada observação.

As técnicas de deteção de outliers são baseadas em pontuações permitindo considerar um limite específico do domínio para selecionar os valores mais relevantes. Técnicas que fornecem classificações binárias para as observações não permitem que uma tal escolha seja possível, embora isso possa ser controlado indiretamente através de opções dos parâmetros dentro de cada técnica.

3.3 Deteção não supervisionada de *outliers*

Os métodos de deteção não supervisionada de outliers não exigem dados de treino e, portanto, são mais aplicáveis. As técnicas desta categoria fazem a suposição implícita de que as observações normais são muito mais frequentes do que os outliers. Caso se verifique a situação contrária estas técnicas não são muito bem-sucedidas, isto é, vai classificar uma observação como sendo outlier em vez de 'normal', ou vice-versa.

Alguns métodos de deteção de valores atípicos assumem uma distribuição para os dados, e classificam como outlier qualquer observação que se desvie da distribuição. Outra estratégia comum na deteção de outliers é assumir uma métrica sobre o espaço de variáveis e usar a noção de distância para definir como outlier uma observação que esteja 'longe' das restantes observações.

A partir das descrições acima, observa-se que existem fortes relações entre clustering e deteção de outliers, isto é particularmente verdade em metodologias baseadas na noção de distância entre observações.

Segundo Torgo, em 2010, "outliers são, por definição, casos muitos diferentes e, portanto, eles não devem ser agrupados em grupos com outras observações, porque os valores atípicos estão muito distantes das restantes observações. Isso significa que um bom agrupamento num conjunto de dados não deve incluir outliers em grandes grupos de dados. No máximo, pode-se esperar que os outliers sejam semelhantes a outros outliers, mas por definição são observações raras e, portanto, não devem formar grandes grupos" [36].

Existem três tipos de deteção não supervisionada de valores extremos que estão relacionados com o estudo [7]: a deteção de outliers baseada na estatística, a deteção de outliers baseada no clustering e a deteção de outliers baseada na proximidade.

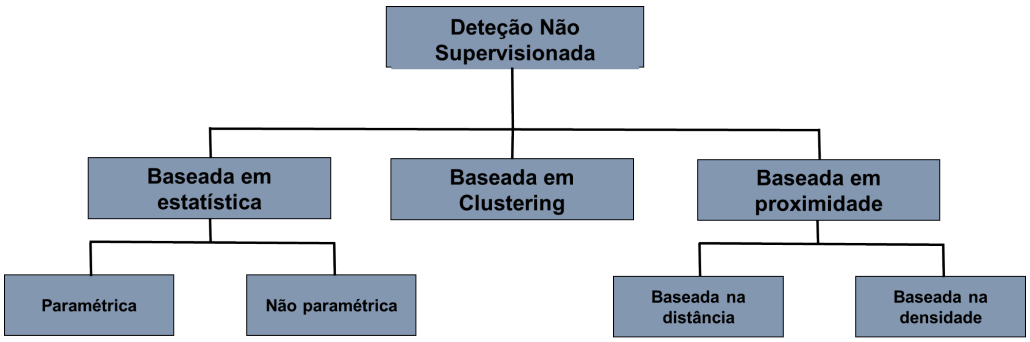

Figura 3.4: Técnicas de deteção não supervisionada de outliers.

3.3.1 Técnicas de deteção de *outliers* baseadas em métodos estatísticos

Os métodos estatísticos de deteção de outliers baseiam-se na seguinte suposição proposta por Anscombe e Guttman, em 1960: uma anomalia é uma observação suspeita de ser parcial ou totalmente irrelevante porque não é gerada por um modelo assumido como estocástico [7].

As técnicas estatísticas introduzem um modelo estatístico (normalmente para o comportamento normal, e que pode ser paramétrico ou não paramétrico) para os dados e, em seguida, aplica-se um teste de inferência estatística para determinar se uma observação pertence ao modelo ou não. As observações que têm uma baixa probabilidade de serem geradas a partir do modelo, com base na estatística de teste, são denominadas de outliers.

Sempre que as suposições do modelo estatístico forem verdadeiras, estas técnicas fornecem uma solução justificável para a deteção de valores atípicos e a probabilidade de uma observação ser um outlier está associada a um intervalo de confiança. Contudo, a metodologia exibe

também desvantagens: base de dados volumosos nem sempre seguem um modelo estatístico; escolher estatística para o teste de hipóteses não é simples; detetar interações entre atributos nem sempre é possível e estimar os parâmetros para alguns modelos estatísticos é difícil.

Técnicas paramétricas

As técnicas paramétricas assumem que os dados 'normais' (não outliers) são gerados por uma distribuição paramétrica, digamos com função de densidade de probabilidade $f(x, \Theta)$, onde x representa uma observação e Θ o vetor de parâmetros (usualmente estimado a partir dos dados). O *score* da observação da base de dados, x, é o inverso da função de densidade de probabilidade, $f(x, \Theta)$.

Uma outra forma de detetar valores atípicos é aplicando testes de hipóteses estatísticos. Em tais testes, a hipótese nula (H_0) afirma que x é gerado por uma distribuição estimada (com o parâmetro Θ) enquanto que se o teste de hipótese estatístico rejeita H_0, x é chamada de outlier. A estatística de teste do teste de hipóteses pode ser usada para obtenção de um score.

Os métodos paramétricos podem ser classificados num dos seguintes tipos, de acordo com a distribuição assumida [21]: baseado no modelo Gaussiano, baseado no modelo de regressão ou baseado na mistura de distribuições.

Técnicas não paramétricas

A deteção de outliers também pode ser dada por meio de técnicas não paramétricas que utilizam modelos estatísticos não paramétricos, isto é, a estrutura do modelo não é definida *a priori*, mas é determinada a partir de dados fornecidos. Tais técnicas, frequentemente, fazem menos suposições sobre os dados e, portanto, podem ser aplicáveis em mais cenários [7].

Baseada no boxplot

Uma técnica simples, e muito comum, na deteção de outlier, consiste em declarar que todas as observações da base de dados que estão a mais de 3σ de distância da média são outliers.

Nota: se conhecer a média populacional, pode-se usar μ (isso quase nunca acontece); se não conhecer, terá de ser $\bar{x}$, a média amostral.

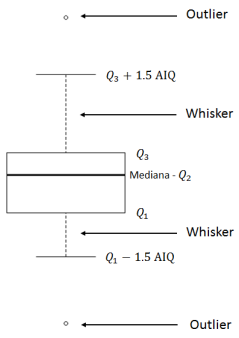

Figura 3.5: Representação do boxplot.

O desvio-quartil ou amplitude inter-quartil (AIQ) é a diferença entre o terceiro e o primeiro quartil, $AIQ = Q_3 - Q_1$, que é uma medida de dispersão que não é afetada por valores extremos. A amplitude inter-quartil é utilizada para a construção do boxplot (figura 3.5Representação do boxplot. figure.caption.26). O boxplot permite a identificação de outliers. Esta técnica é um dos métodos mais simples da aplicação de técnicas estatísticas para a deteção univariada e multivariada de outliers [32].

O boxplot permite representar a mediana Q_2, o quartil inferior Q_1, o quartil superior Q_3 e a amplitude inter-quartil AIQ. A linha central da caixa que constitui o boxplot é referente a mediana do conjunto de dados, como se pode observar a caixa é delimitada pela linha superior Q_3 e a inferior Q_1.

O segmento de reta vertical conecta o topo da caixa ao maior valor observado e o outro segmento conecta a base da caixa ao menor valor observado, este segmento denomina-se por Whisker ou fio de bigode. A haste inferior do fio de bigode é desde o quartil inferior até ao menor valor não inferior, $Q_1 - 1.5AIQ$, e a haste superior do fio de bigode é desde o quartil superior até ao maior valor não superior, $Q_3 + 1.5AIQ$. Os valores superiores a $Q_3 + 1.5AIQ$ e inferiores a $Q_1 - 1.5AIQ$ são denominados de outliers.

O boxplot permite avaliar a simetria dos dados, a dispersão e a existência de outliers.

Baseada no histograma

Técnicas baseadas em histogramas são particularmente populares na comunidade de deteção de intrusão e fraudes.

A técnica de deteção de outliers baseada no histograma é formada por duas etapas. A primeira etapa consiste na **construção do histograma** usando a base de dados amostral. Note-se que, embora os métodos não paramétricos não assumam qualquer modelo estatístico *a priori*, frequentemente, requerem parâmetros especificados pelo utilizador para ajustar modelos a partir dos dados. Por exemplo, para construir um histograma, o utilizador precisa especi-

ficar o tipo de histograma (largura) e outros parâmetros (o número de barras ou o tamanho das barras). A segunda etapa consiste na **deteção dos outliers**, a técnica verifica se uma observação pertence a qualquer uma das barras do histograma. Em caso positivo, a observação é considerada 'normal', caso contrário é um outlier.

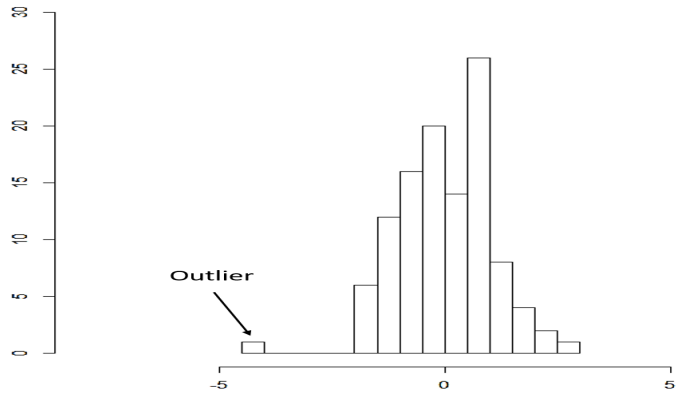

Figura 3.6: Representação da técnica baseada no histograma.

Como se pode verificar no histograma, figura 3.6Representação da técnica baseada no histograma. figure.caption.27, tem-se uma das barras mais distante das restantes, o que indica ser um possível outlier.

Uma desvantagem em usar histogramas como um modelo não paramétrico para a deteção de outliers é o facto de ser difícil escolher um tamanho do compartimento apropriado. Por um lado, se o tamanho da barra estiver muito pequena, muitas observações 'normais' podem acabar em barras vazias ou raras e, assim, ser identificados indevidamente como outliers. Isto leva a uma elevada taxa de falsos positivos e a uma baixa precisão. Por outro lado, se o tamanho da barra estiver muito alta, as observações atípicas podem-se infiltrar em algumas barras frequentes e, assim, ser classificadas como 'normais'. Isto leva a uma elevada taxa de falsos negativos.

3.3.2 Técnicas de deteção de *outliers* baseadas em clustering

A noção de outliers está relacionada com a de clusters. O clustering, como já foi referido na secção anterior, 'compara' as observações, e agrupa no mesmo cluster as observações mais semelhantes e as mais distintas ficam em clusters diferentes. Intuitivamente, um outlier é uma observação que pertence a um cluster pequeno e remoto, ou não pertence a nenhum cluster, pois são observações atípicas. Isso leva a três categorias gerais para a deteção de outliers baseadas em métodos de clustering [16].

Primeira categoria

Na primeira categoria, as técnicas que são baseadas nesta suposição aplicam um algoritmo que não obriga a cada observação da base de dados pertencer a um cluster. As observações que não pertencem a nenhum cluster são consideradas outliers. Uma série de técnicas de deteção de outliers seguem esta abordagem, tal como o **DBSCAN** [12].

Segunda categoria

A segunda categoria supõe que as observações 'normais' da base de dados encontram-se próximas do centróide do respetivo cluster, enquanto que as observações denominadas de outliers estão muito distantes do centróide do cluster mais próximo. Técnicas baseadas nesta suposição consistem em duas etapas: os dados são agrupados usando um algoritmo de clustering e para cada observação da base de dados é determinada a distância ao centróide mais próximo, sendo essa distância o valor do score de outliers.

Uma série de técnicas de deteção de outliers seguem esta abordagem, tais como mapa auto-organizável (SOM) e o método de K-médias (já referido na secção anterior).

Terceira categoria

Na terceira categoria, as observações da base de dados 'normais' formam clusters grandes e densos, enquanto que os outliers formam clusters pequenos ou dispersos. As observações que pertencem a clusters cujo tamanho e / ou densidade estão abaixo do limite do score de outlier são denominadas de outliers.

He et al., em 2003, propôs uma nova definição para outlier: **outlier local baseado em clusters**. Para identificar o significado 'físico' da definição do outlier, atribui-se a cada observação um fator de outlier, **CBLOF**, que é medido pelo tamanho do cluster ao qual a observação pertence, e a distância entre as observações e o cluster mais próximo (se a observação está num cluster pequeno).

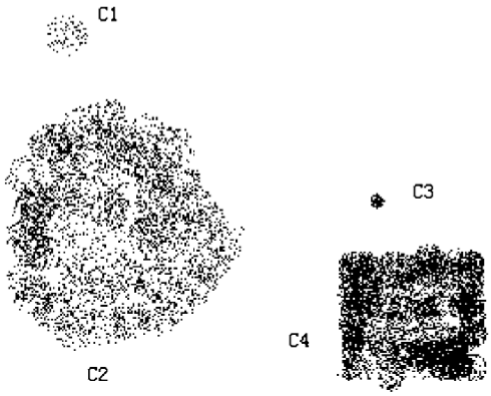

Figura 3.7: Representação do CBLOF de um conjunto de dados bidimensional. [19]

Para uma melhor compreensão da técnica CBLOF considera-se a figura 3.7Representação do CBLOF de um conjunto de dados bidimensional. [19] figure.caption.31, em que é possível observar um conjunto de dados composto por quatro clusters, C_1, C_2, C_3 e C_4. He et al. referem que as observações dos clusters C_1 e C_3 devem ser consideradas outliers, uma vez que não estão contidas em nenhum dos grandes clusters. Esse será precisamente o resultado da aplicação da técnica Factor de Outlier Local Baseado em Clusters [19].

Com o fator de outlier, CBLOF, podemos determinar o grau de desvio de uma observação, definição 1Terceira categoriadefi.1.

Definição 1 *(Factor de Outlier Local Baseado em Clusters):*
Suponha-se que $C = \{c_1, c_2, ..., c_k\}$ *é o conjunto de clusters, na seguinte sequência* $\mid c_1 \mid \geq \mid c_2 \mid \geq ... \geq \mid c_k \mid$. *Para qualquer observação t, o* $CBLOF(t)$ *é definido por:*

$$CBLOF(t) = \begin{cases} \mid c_i \mid \times min(dist(t, c_j)), \ \ onde\ t \in c_i, c_i \in SC\ e\ c_j \in LC,\ para\ j = \{1, ..., b\}\ ; \\ \mid c_i \mid \times (dist(t, c_i)),\ onde\ t \in c_i \ \ e\ c_i \in LC \end{cases} \tag{3.1}$$

onde b é definido como o limite de clusters de grandes dimensões e pequenas dimensões se uma das seguintes fórmulas é válida:

$$(\mid c_1 \mid \geq \mid c_2 \mid \geq ... \geq \mid c_k \mid) \geq \mid D \mid \times \alpha; e \tag{3.2}$$

$$\frac{\mid c_b \mid}{\mid c_{b+1} \mid} \geq \beta, \tag{3.3}$$

onde α *e* β *são dois parâmetros numéricos. Então, o conjunto de cluster grande é definido por:* $LC = \{c_i, \mid i \leq b\}$ *e o conjunto de cluster pequeno é definido por:* $SC = \{c_j, \mid j > b\}$.

Observa-se que se $\alpha = 90\%$, os clusters contêm 90% dos pontos da base de dados como sendo clusters grandes e se definirmos $\beta = 5$, o tamanho de qualquer cluster em LC é de pelo menos cinco vezes maior do que os clusters em SC.

Para calcular o $CBLOF(t)$, precisamos primeiro de um algoritmo de clustering que produza bons resultados de agrupamento. Para tal, aplica-se o algoritmo FindCBLOF. Este algoritmo primeiro particiona o conjunto de dados em clusters com um algoritmo de clustering. Os clusters grandes e os pequenos são derivados usando os parâmetros α e β. Em seguida, para cada ponto da base de dados, o valor de CBLOF é calculado. As observações cujo valor CBLOF seja elevado são outliers.

Input:

D: base de dados que contém n observações;

α e β: dois parâmetros numéricos; e

A: algoritmo de clustering.

Output:

O conjunto de outliers.

O algoritmo FindCBLOF é da seguinte forma:

1. Particiona a base de dados num conjunto de K clusters usando o algoritmo de clustering, A, assim $C = A(D, K, \zeta)$, onde $C = \{c_i, i = 0, ..., K-1\}$ e ζ é o conjunto de parâmetros do algoritmo de clustering, A. Este passo deve retorna algo desta forma: $\mid c_1 \mid \geq \mid c_2 \mid \geq ... \geq \mid c_k \mid$;

2. Obtém-se o LC e o SC usando os parâmetros α e β;

3. Para cada observação t do conjunto de dados D é calculado o valor do CBLOF, da seguinte forma:
   Se $t \in c_i$ e $c_i \in SC$, então $CBLOF(t) = \mid c_i \mid \times min(dist(t, c_j)),$ onde $t \in c_i,$ e $c_j \in LC$;
   Caso contrário, $CBLOF(t) = \mid c_i \mid \times (dist(t, c_i)),$ onde $t \in c_i$ e $c_i \in LC$;

4. É devolvida uma lista de observações, outliers, cujo valor CBLOF é o mais alto.

A eficiência do algoritmo FindCBLOF na deteção de outliers em base de dados é limitada, devido à qualidade da técnica de clustering aplicada.

3.3.3 Técnicas de deteção de *outliers* baseadas na proximidade

O conceito de análise do vizinho mais próximo tem sido utilizado em várias técnicas de deteção de outliers. Chandola et al. supuseram que tais técnicas determinam que "as observações 'normais' da base de dados ocorrem em espaços densos, enquanto as observações atípicas, outliers, ocorrem longe dos seus vizinhos mais próximos" [8].

As técnicas de deteção de outliers baseadas na proximidade exigem uma medida de distância ou semelhança entre duas observações. A distância (ou semelhança) entre duas observações pode ser calculada de diferentes formas.

Para dados contínuos, a distância de Minkowski é geralmente utilizada para calcular a distância entre dois pontos $\mathbb{R}^n(n > 1)$. Em particular, a distância de Minkowski de ordem 1 (Manhattan) e ordem 2 (Euclidiana) são as duas medidas de distância mais utilizadas.

A noção de distância para dados categóricos não é tão direta quanto para dados contínuos. A característica dos dados categóricos é que os diferentes valores que um atributo toma não são inerentemente ordenados. Assim, não é possível comparar diretamente dois valores categóricos diferentes. A maneira mais simples de encontrar semelhança entre dois atributos é atribuir 1 se os valores forem idênticos e no caso de não serem semelhantes atribuir o valor de 0. Para dois pontos categóricos multivariados, a semelhança entre eles será diretamente proporcional ao seu número de atributos.

A maioria das técnicas que serão discutidas, bem como as técnicas baseadas em clusters não exigem que a medida de distância seja estritamente métrica. Estas técnicas não exigem que as medidas de distâncias satisfaça a propriedade da desigualdade triangular.

As técnicas de deteção de outliers baseadas na proximidade podem ser agrupadas em duas categorias: técnicas que usam a distância, entre uma observação e os seus K vizinhos mais próximos, como sendo o score de outlier, e técnicas que determinam a densidade relativa de cada observação da base de dados, para calcular o score de outlier.

Técnicas baseadas na distância

Métodos baseados na distância são uma classe popular das técnicas de deteção de outliers, pois determinam o score de outlier de uma observação, com base nas distâncias dos K vizinhos mais próximos. Estes métodos funcionam com a suposição de que as distâncias dos K vizinhos mais próximos dos pontos outliers são muito maiores do que a dos pontos 'normais'. Diferentes variações desta definição especificam K como um número absoluto de observações mais próximas (K vizinhos mais próximos) de um determinado ponto.

Os métodos baseados na distância podem possibilitar uma melhor capacidade de distinguir entre valores fracos e fortes em conjuntos de dados ruidosos, comparando com os métodos de clustering. Como se pode observar no caso da figura 3.8Algoritmos baseados nos K vizinhos mais próximos podem ser mais eficazes do que algoritmos baseados em clustering em bases de dados com muito ruído. [1] figure.caption.33, um algoritmo baseado em clustering não será capaz de distinguir facilmente entre ruído e outlier. Isso ocorre porque a distância do ponto A até ao

centróide do cluster mais próximo permanecerá a mesma nas figuras (a) e (b). Por outro lado, as técnicas baseadas nos K vizinhos mais próximos distingue muito melhor porque os pontos ruidosos serão incluídos entre as avaliações da distância, ao invés dos centróides dos clusters. Na figura (a) observam-se claramente dois clusters distintos e um outlier, A, enquanto que na figura (b) a definição dos clusters é mais complicada pois existem muitos pontos ruidosos [1].

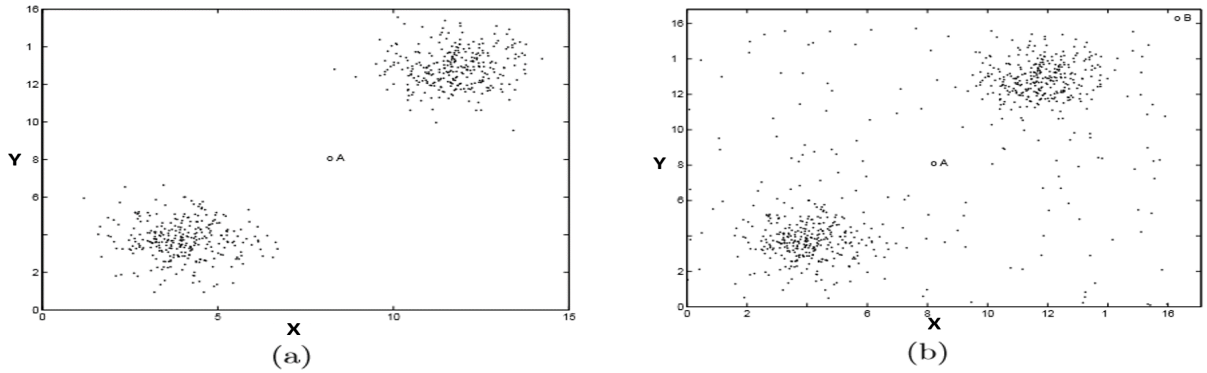

Figura 3.8: Algoritmos baseados nos K vizinhos mais próximos podem ser mais eficazes do que algoritmos baseados em clustering em bases de dados com muito ruído. [1]

Nos métodos baseados na distância um conjunto isolado de outliers estreitamente relacionados os pontos 'anormais' também podem ser identificados devido ao valor apropriado de K. Enquanto que nos métodos baseados em clustering, esses pontos podem enviesar as representações dos clusters.

Um dos primeiros estudos sobre a deteção de outliers baseado em métodos de distância foi devido a Knorr et al., em 1998, no qual definem outlier da seguinte forma: "Uma observação x, num conjunto de dados D é um outlier, $DB(p, d)$, se pelo menos a fração p das observações em D estão a uma distância maior que d de x" [20]. Note-se que o termo $DB(p, d)$-outlier usa-se como nota abreviada para um outlier baseado em métodos de distância. Os valores de p e d fornecem indicações sobre tão 'forte' é um outlier identificado. Muitas abordagens existentes para encontrar outliers não fornecem tais indicações.

Para um conjunto de dados, D, de n observações a serem analisadas, um usuário pode especificar um limite de distância, d, para definir uma vizinhança razoável de uma observação. Para cada observação, x, pode-se examinar o número de outras observações na vizinhança d de x. Se a maioria das observações em D estão longe de x, ou seja, não na d-vizinhança de x, então x pode ser considerado como um outlier. Formalmente, d ($d > 0$), é o limite de distância e p ($0 < p \leq 1$) é o limite de fração. Uma observação, x, é um $DB(p, d)$-outlier se:

$$\frac{\parallel \{x' \mid dist(x, x') \leq d\} \parallel}{\parallel D \parallel} \leq p, \tag{3.4}$$

onde $dist(\cdot,\cdot)$ é uma medida de distância e $x' \in D$. Equivalentemente, pode-se determinar se uma observação, x, é um $DB(p,d)$-outlier, verificando a distância entre x e os seus K vizinhos mais próximos, x_K, onde $K = \lceil p \parallel D \parallel \rfloor$. A observação, x, é um outlier se $dist(x, x_K) > d$, pois há menos de K observações que estão na d-vizinhança de x.

Input:

D: base de dados que contém n observações;

x_i: observações da base de dados, $x_i = \{x_1, \ldots, x_n\}$;

d: parâmetro da distância, $(d > 0)$; e

p: parâmetro da fração das observações em D, $(0 < p \leq 1)$.

Output:

$DB(p,d) > d$ - outlier em D.

O algoritmo de deteção de outliers baseado na distância é da seguinte forma:

1. Seleciona uma observação, aleatória, $x_i, i = 1, ..., n$;
2. Inicia a contagem das observações vizinhas, $conta = 0$;
3. Seleciona uma observação, $x_j, j = 1, ...n$;
4. Se $x_j \neq x_i$ e a $dist(x_i, x_j) \leq d$ então conta = conta + 1;
   - Se conta $\geq p \times n$ então devolva a informação

     $$\{x_i \text{ não pode ser } DB(p,d) - \text{outlier}\};$$

   Caso contrário, devolve

   $$\{x_i \text{ é um } DB(p,d) - \text{outlier}\}; \text{e}$$

5. Continua o processo até que as n observações tenham todas sido processadas.

Algoritmo de deteção de outliers baseado na distância:

- **KNN** (K vizinhos mais próximos)
  KNN é um método proposto por Ramaswamy et al., em 2000, que deteta outliers através da distância entre uma observação x e o seu K-ésimo vizinho mais próximo e como resultado apresenta o top-n de outliers [28]. Este algoritmo não exige que o usuário especifique o parâmetro de distância, d. Em vez disso, baseia-se na distância K-*ésimo* vizinho mais próximo de um ponto. Para qualquer inteiro positivo K e um ponto p,

obtém-se $K-$distância(p), isto é, a distância do K-*ésimo* vizinho mais próximo de p. Por exemplo, os pontos com valores elevados de $K-$distância(p) têm vizinhanças mais esparsa e, portanto, são geralmente mais fortes do que os pontos pertencentes a clusters densos que tendem a ter valores mais baixos de $K-$distância(p).

Num contexto de KNN, um outlier é definido da seguinte forma: "Seja D um conjunto de dados com N observações. Uma observação $p \in D$ é considerada um outlier se não houver mais do que $n-1$ elementos p' de tal modo que $K-$distância$(p') > K-$distância(p), onde n e K são dois parâmetros" [28]. Por outras palavras, se classificarmos as observações de acordo com a distância, $K-$distância(p), as top-n observações desse ranking são consideradas outliers. O mais importante para o utilizador é obter os top-n outliers.

Com a definição adquirida para outliers, é possível classificar outliers com base nas distâncias $K-$distância(p). Os outliers são os pontos cujo valor $K-$distância(p) é maior.

Técnicas baseadas na densidade

Relembrando a definição de outlier, por Knorr et al., em 1998, em que uma observação, x, numa base de dados, D, é um $DB(p,d)$-outlier se pelo menos a percentagem, $pct = p \times 100\%$, das observações em D estão a uma distância maior que d de x, ou seja, a cardinalidade do conjunto $\{x' \in D \mid dist(x,x') \leq d\}$ é menor ou igual a $(100 - pct)\%$ do tamanho de D, onde $dist(x,x') = min\{dist(x,x') \mid x' \in C\}$, sendo C o cluster. Nesta definição apenas algumas das observações são detetadas como sendo outliers, uma vez que a definição tem uma visão 'global' do conjunto de dados. Esses outliers podem ser vistos como outliers 'globais' [20].

Para uma observação 'normal' situada numa região densa, a sua densidade local será semelhante à dos seus vizinhos, enquanto que para uma observação outlier, a sua densidade local será menor que a dos seus vizinhos. Estes outliers são considerados outliers 'locais'. Para uma melhor compreensão, considera-se o exemplo da figura 3.9Representação de uma base de dados bidimensional. [6] figure.caption.35, onde é representado um conjunto da dados bidimensional, contendo 502 observações, das quais 400 pertencem ao cluster C_1, 100 ao cluster C_2 e duas observações adicionais O_1 e O_2. Neste exemplo pode-se observar que o cluster C_2 comparado com o cluster C_1 é mais denso. Com a noção de outlier 'local' os pontos O_1 e O_2 são considerados outliers. Em contraste com a visão global do conjunto de dados, usando a deteção de outliers baseado na distância, apenas se consideraria O_1 como outlier, pois O_2 só seria considerado outlier se o limite de distância fosse menor. No entanto, ao alterar-se a distância para uma menor, resultaria que muitos dos pontos do cluster C_1 podem ser incorretamente detetados

como outliers. Isto também significa que o score de outlier retornado por um algoritmo baseado na distância para a deteção de outliers é incorreto quando houver heterogeneidade significativa nas distribuições locais dos dados [6].

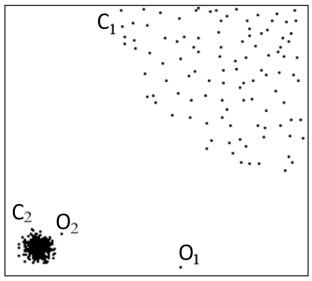

Figura 3.9: Representação de uma base de dados bidimensional. [6]

O exemplo acima mostra que a visão global tomada por $DB(pct, d)$-outlier é adequada sob certas condições mas não satisfatória para o caso geral quando existem grupos de diferentes densidades.

Algoritmos de deteção de outliers baseados na densidade:

- **Fator de Outlier Local** (LOF):
  Breunig et al., em 2000, começa por incluir as definições 2Técnicas baseadas na densidadedefi.2 e 3Técnicas baseadas na densidadedefi.3, de forma a se ter noção da vizinhança $K-$ distância de p, e, correspondentemente, da distância de alcance de uma observação p em relação a uma observação O [6].

  Definição 2 *(A vizinhança K - distância de uma observação p)*
  Seja a $K-$distância de p, a vizinhança K - distância de p que contém todas as observações cuja distância para p não é maior que $K-$ distância , ou seja,

$$N_{K-distância(p)}(p) = \{q \in D \setminus \{p\} \mid dist(p,q) \leq K - distância(p)\},$$

  onde q são chamadas de K - vizinhos mais próximos de p.

  Definição 3 *(A distância de alcance de uma observação p em relação a uma observação O)*
  Seja K um número natural fixo. A distância de alcance da observação p em relação à observação O é definida por:

$$distância\text{ - }alcance_K(p,O) = max\{K - distância(O), dist(p,O)\}. \tag{3.5}$$

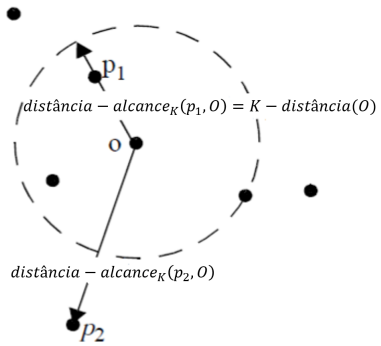

Figura 3.10: Representação distância - alcance$_K(p_1, O)$ e distância - alcance$_K(p_2, O)$, para $K = 4$. [6]

Para uma melhor compreensão da definição 3Técnicas baseadas na densidadedefi.3, inclui-se a figura 3.10Representação distância - alcance$_K(p_1, O)$ e distância - alcance$_K(p_2, O)$, para $K = 4$. [6] figure.caption.36 que ilustra a ideia da distância de alcance. O ponto p_2 está distante do ponto O, então a distância de alcance entre os dois é a distância real entre eles, ou seja, $dist(p_2, O)$. No entanto, no caso do ponto p_1, ele está suficientemente próximo do ponto O; neste caso, a distância real é substituída pela $K-$ distância(O).

Note-se que quanto maior o valor de K, mais semelhantes são as distâncias de alcance para observações dentro da mesma vizinhança.

Num algoritmo típico de clustering baseado na densidade existem dois parâmetros que definem a noção de densidade: um parâmetro 'MinPts' que especifica o número mínimo de observações e um outro que especifica o volume.

Esses dois parâmetros determinam um limite de densidade para aplicar aos algoritmos de clustering, ou seja, observações ou regiões estão conectadas se as densidades da vizinhança excederem o limite de densidade. Para detetar os outliers baseados na densidade é necessário comparar as densidades dos diferentes conjuntos de observações. Portanto, mantém-se 'MinPts' como o único parâmetro e usam-se os valores da distância-alcance$_{\text{MinPts}}$ (p, O), para $O \in N_{\text{MinPts}}(p)$, como medida do volume para determinar a densidade na vizinhança da observação p.

Definição 4 *(Densidade de alcance local de uma observação p)*
A densidade de alcance local de p é dada pela seguinte expressão:

$$dal_{MinPts}(p) = 1 \setminus \left(\frac{\sum_{O \in N_{MinPts}(p)} distância\text{-}alcance_{MinPts}(p, O)}{\mid N_{MinPts}(p) \mid} \right)$$
$$= \frac{\mid N_{MinPts}(p) \mid}{\sum_{O \in N_{MinPts}(p)} distância\text{-}alcance_{MinPts}(p, O)}.$$

Intuitivamente, como se pode observar pela a expressão da definição 4Técnicas baseadas na densidadedefi.4, a densidade de alcance local de uma observação p é o inverso da distância média de alcance com base nos vizinhos mais próximos de p. Note-se que a densidade local pode ser ∞ se a soma de todas as distâncias de alcance forem 0, isso pode ocorrer quando existe pelo menos 'MinPts' iguais a p na base de dados.

Definição 5 *(LOF de uma observação p)*
O fator de outlier local de p é definido por:

$$LOF_{MinPts}(p) = \frac{\sum_{O \in N_{MinPts}(p)} \frac{dal_{MinPts}(O)}{dal_{MinPts}(p)}}{\mid N_{MinPts}(p) \mid} = \sum_{O \in N_{MinPts}(p)} dal_{MinPts}(O) \cdot \sum_{O \in N_{MinPts}(p)} distância\text{-}alcance_{MinPts}(p, O) \cdot$$

O fator de outlier da observação p retorna o grau que se denomina p um outlier, que consiste na média da razão entre a densidade de alcance local de p e 'MinPts'- vizinhos mais próximos de p. Conclui-se, pela expressão da definição 5Técnicas baseadas na densidadedefi.5, que quanto mais baixo for o valor da densidade de alcance local de p e quanto mais alto for o valor da densidade de alcance local dos 'MinPts'-vizinhos mais próximos de p, maior será o valor do LOF de p.

Input:
D: base de dados que contém n observações; e
K: número mínimo de observações.

Output:
O conjunto dos top-n outliers.

O algoritmo para calcular o valor de LOF de um conjunto de dados é o seguinte:

1. Seleciona um ponto p, $p \in D$, e O um vizinho de p;
2. Calcula todas as $dist(p, O)$;
3. Calcula todas as $K - \text{distância}(p)$;
4. Calcula todas as $N_{K-\text{distância}(p)}(p)$;
5. Calcula todas as $dal_K(p)$;
6. Calcula todos os valores de $LOF_K(p)$;
7. Ordena de forma decrescente todos os valores do $LOF_K(p)$;

8. Retorna os top-n outliers.

Para uma melhor compreensão do algoritmo, que permite calcular o valor LOF, apresenta-se um exemplo em anexo AExemplo da aplicação do algoritmo LOFAnexo.a.A que o explicita.

O teorema 1Técnicas baseadas na densidadeteo.1 que se segue devolve um majorante e um minorante para o valor de LOF para qualquer observação p. Para qualquer observação p, a distância-alcance$_{\min}(p)$ representa a distância mínima de alcance entre p e um 'MinPts'-vizinho mais próximo de p, ou seja,

$$\text{distância-alcance}_{\min}(p) = min\{\text{distância-alcance}(p, q) \mid q \in N_{\text{MinPts}}(p)\}.$$

Da mesma forma, distância-alcance$_{\max}(p)$ denota a distância máxima de alcance, ou seja,

$$\text{distância-alcance}_{\max}(p) = max\{\text{distância-alcance}(p, q) \mid q \in N_{\text{MinPts}}(p)\}.$$

Além disso, para generalizar essas definições para o 'MinPts'-vizinho mais próximo q de p, onde in-distância-alcance$_{\min}(p)$ refere-se à distância mínima de alcance entre q e um 'MinPts'-vizinho mais próximo de q, isto é,

$$\text{in-distância-alcance}_{\min}(p) = min\{\text{distância-alcance}(q, O) \mid q \in N_{\text{MinPts}}(p) \text{ e } O \in N_{\text{MinPts}}(q)\}.$$

Da mesma forma, in-distância-alcance$_{\max}(p)$ denota o máximo. Os autores referem 'MinPts'-vizinhos mais próximos de p como sendo a distância-alcance da vizinhança de p e referem 'MinPts'- vizinhos mais próximos de q como sendo a in-distância-alcance da vizinhança de p, sempre que q é um 'MinPts'-vizinho mais próximo de p.

Teorema 1 *Seja p uma observação da base de dados D e* $1 \leq$ *'MinPts'* $\leq \mid D \mid$. *Tem-se que:*

$$\frac{\textit{distância-alcance}_{min}(p)}{\textit{in-distância-alcance}_{max}(p)} \leq LOF(p) \leq \frac{\textit{distância-alcance}_{max}(p)}{\textit{in-distância-alcance}_{min}(p)} \tag{3.6}$$

A figura 3.11Ilustração do teorema 1Técnicas baseadas na densidadeteo.1. [6] figure.caption.37 ilustra um exemplo simples das definições anteriores.

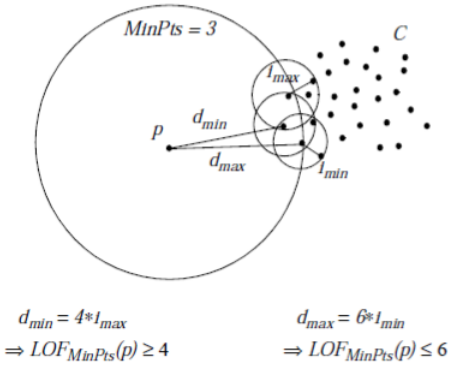

Figura 3.11: Ilustração do teorema 1Técnicas baseadas na densidadeteo.1. [6]

O valor de LOF aumenta quando o valor da distância-alcance é menor que o valor in-distância-alcance. Assim quando se obtém o valor $LOF(p) \simeq 1$ conclui-se que a observação p não é um outlier, pois a distância de p a q, sendo q o vizinho mais próximo de p, é igual à distância de q a um vizinho mais próximo deste. A distância-alcance$_{\min}(p)$ é representado na figura 3.11Ilustração do teorema 1Técnicas baseadas na densidadeteo.1. [6] figure.caption.37 por $d_{\min}$ e o valor distância-alcance$_{\max}(p)$ é denotado por $d_{\max}$. Como p está relativamente distante de C e considerando 'MinPts' $= 3$, a $3-$distância de cada observação q em C é muito menor do que a distância real entre p e q. Assim, a partir da definição 3Técnicas baseadas na densidadedefi.3, a distância de alcance de p em relação a q é dada pela distância real entre p e q. Considerando os 3-vizinhos mais próximos de p é possível achar a distância mínima e a máxima de alcance aos seus 3-vizinhos mais próximos. Na figura 3.11Ilustração do teorema 1Técnicas baseadas na densidadeteo.1. [6] figure.caption.37, a in-distância-alcance$_{\min}(p)$ e a in-distância-alcance$_{\max}(p)$ são denotadas por $i_{\min}$ e $i_{\max}$, respetivamente.

Para ilustrar o teorema usando o exemplo da figura 3.11Ilustração do teorema 1Técnicas baseadas na densidadeteo.1. [6] figure.caption.37, d_{min} é 4 vezes maior do que i_{max} e d_{max} é 6 vezes maior que i_{min}. Pelo teorema, conclui-se que o valor de LOF de p está entre 4 e 6, $4 \leq LOF(p) \leq 6$. É de salientar que para observações próximas num cluster, o valor de LOF dessas observações é próximo de 1.

Note-se que o valor de LOF pode variar dependendo do 'MinPts'. Breunig et al. fornecem indicações sobre como este intervalo de valores 'MinPts' pode ser escolhido. Seja 'MinPtsLB' e 'MinPtsUB', o "limite inferior" e o "limite superior", respetivamente, do intervalo.

O valor 'MinPtsLB' mínimo que se considera é 2. No entanto, é aconselhável remover flu-

tuações estatísticas indesejadas. Por outro lado, para escolher o melhor valor 'MinPtsLB' considera-se uma observação p e um cluster C. Se C contém menos do que 'MinPtsLB' observações, o conjunto de 'MinPts'-vizinhos mais próximos de cada observação em C incluirá p, e vice-versa. Assim, aplicando o teorema 1Técnicas baseadas na densidadeteo.1, o LOF de p e todas as observações em C serão bastantes semelhantes, tornando assim p indistinguível das observações em C. Se, por outro lado, C contiver mais do que 'MinPtsLB' observações, os 'MinPts' - vizinhos mais próximo das observações em C não irão conter p, mas algumas observações de C serão incluídas na vizinhança de p. Assim, dependendo da distância entre p e C e a densidade de C, o LOF de p pode ser bastante diferente do de uma observação em C. É de salientar que o 'MinPtsLB' pode ser considerado como o número mínimo de observações que um cluster tem que conter, de modo a que as restantes observações possam ser outliers locais relativamente a esse cluster. Esse valor depende da base de dados em estudo.

Seja C um conjunto de observações 'próximas'. Então 'MinPtsUB' pode ser considerado como a cardinalidade máxima de C de todas as observações que podem ser potenciais outliers locais. Por 'próxima' quer-se dizer que os valores da distância-alcance$_{\min}$, distância-alcance$_{\max}$, in-distância-alcance$_{\min}$ e in-distância-alcance$_{\max}$ são todos muito semelhantes. Neste caso, para valores de 'MinPts' que excedem o 'MinPtsUB', o teorema 1Técnicas baseadas na densidadeteo.1 requer que o valor de LOF de todas as observações em C seja próximo de 1. Portanto, a sugestão que os autores fornecem para escolher o melhor valor de 'MinPtsUB' é o número máximo de observações 'próximas' que podem ser outliers locais.

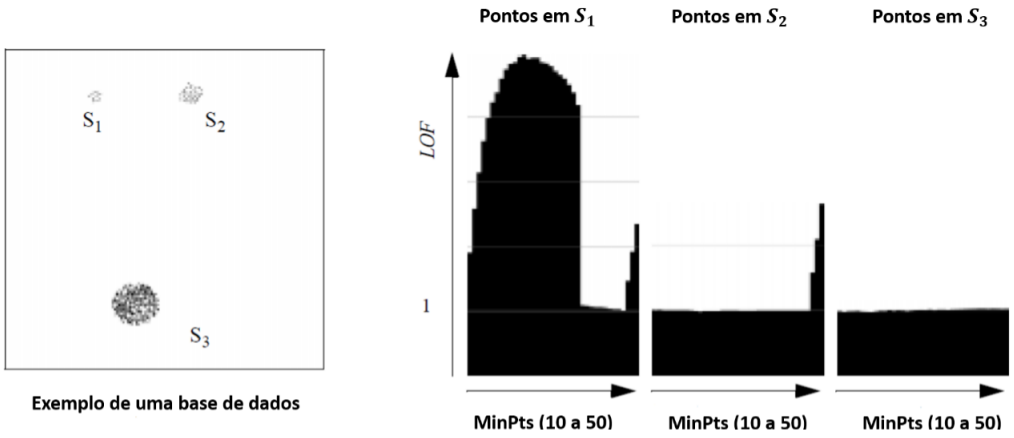

Figura 3.12: Intervalos de valores de LOF para diferentes observações numa base de dados. [6]

Para uma melhor compreensão, considera-se o exemplo da figura 3.12Intervalos de valores

de LOF para diferentes observações numa base de dados. [6] figure.caption.38, onde o conjunto S_1 contém 10 observações, o S_2 contém 35 e o S_3 contém 500 observações. É de notar que no conjunto S_3 nenhuma observação será outlier devido ao valor de LOF ser próximo de 1, no entanto em S_1 as observações são outliers para valores 'MinPts' entre 10 e 35. As observações em S_2 são outliers quando MinPts = 45. A razão para esses efeitos é que, começando por definir MinPts = 36, os 'MinPts' - vizinhos mais próximos das observações em S_2 começam a incluir algumas observações do conjunto S_1, pois S_2 só contém 35 observações. No caso em que MinPts = 45, os membros deste conjunto 'combinado' de observações nos conjuntos S_1 e S_2 começam a incluir observações de S_3 nas vizinhanças e, assim, começam a tornar-se outliers em relação a S_3 [6].

Tendo determinado 'MinPtsLB' e 'MinPtsUB', podemos calcular para cada observação o valor de LOF dentro desse intervalo. O score de uma observação p é baseado em:

$$max\{\text{LOF}_{\text{MinPts}}(p) \mid \text{MinPtsLB} \leq \text{MinPts} \leq \text{MinPtsUB}\}.$$

Dado todos os valores de LOF dentro do intervalo, em vez de se considerar o máximo, pode-se considerar outros agregados, como por exemplo o mínimo ou a média.

- **Integral de Correlação Local** (LOCI):

  O método integral de correlação local (LOCI) foi proposto por Papadimitriou et al., em 2003. LOCI é um método eficaz para a deteção de outliers, que tem algumas vantagens [25]. O próprio método fornece um corte automático, para determinar se uma observação é ou não um outlier, em contraste com os métodos referidos anteriormente em que é necessário o usuário indicar o valor de corte. O método LOCI representa um gráfico para cada ponto, este gráfico resume informações sobre a vizinhança da observação, determinando clusters, seus diâmetros e distâncias entre clusters. A distância aplicada para obter o gráfico é a distância de Jaccard.

  Note-se que o coeficiente de Jaccard, $J(A, B)$, mede a semelhança entre conjuntos da base e é definido como o tamanho da interseção dividido pelo tamanho da união dos conjuntos. A distância de Jaccard, $d_J(A, B)$, mede a semelhança entre conjuntos, é complementar ao coeficiente de Jaccard e é obtida pela seguinte expressão:

  $$d_J(A, B) = 1 - J(A, B) = \frac{\mid A \cup B \mid - \mid A \cap B \mid}{\mid A \cup B \mid}.$$

  O algoritmo LOCI calcula os valores do fator de desvio de multi-granularidade (MDEF) e do σ_{MDEF}, para todas as observações, e em seguida determina um outlier sempre que o MDEF é três vezes mais do que o valor do σ_{MDEF}.

Seja a r-vizinhança de uma observação p_i, o conjunto de observações contidas na distância r de p_i. Intuitivamente, o fator de desvio de multi-granularidade no raio r para uma observação p_i é o desvio relativo da densidade da vizinhança local da densidade média da vizinhança local na sua r-vizinhança. Assim, uma observação cuja densidade da vizinhança corresponde à densidade média da vizinhança local terá um valor de MDEF= 0, enquanto que, uma observação que seja outlier terá o valor de MDEF muito superior a 0.

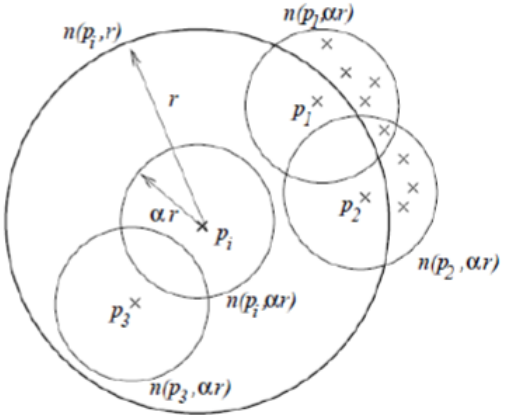

Figura 3.13: Representação da n e $\widehat{n}$ - por exemplo $n(p_i, r) = 4, n(p_r, \alpha r) = 1, n(p_1, \alpha r) = 6$ e $\widehat{n}(p_i, r, \alpha) = (1 + 6 + 5 + 1)/4 = 3.25$. [25]

Seja $n(p_i, \alpha r)$ o número de observações na αr-vizinhança de p_i, isto é, $n(p_i, \alpha r) \equiv |\ N(p_i, r)\ |$ e seja $\widehat{n}(p_i, r, \alpha)$ a média sobre todas as observações p na r-vizinhança de p_i de $n(p_i, \alpha r)$, isto é

$$\widehat{n}(p_i, r, \alpha) \equiv \frac{\sum_{p \in N(p_i, r)} n(p_i, \alpha r)}{n(p_i, r)}, \tag{3.7}$$

como se pode observa na figura 3.13Representação da n e $\widehat{n}$ - por exemplo $n(p_i, r) = 4, n(p_r, \alpha r) = 1, n(p_1, \alpha r) = 6$ e $\widehat{n}(p_i, r, \alpha) = (1+6+5+1)/4 = 3.25$. [25] figure.caption.39.

O uso de dois raios consiste em separar o raio αr do raio r sobre o qual se esta a fazer a média. Denota-se como integral de correlação local a função $\widehat{n}(p_i, \alpha, r)$ sobre todo r.

Definição 6 *Para qualquer p_i, r e α define-se o MDEF no raio r como:*

$$MDEF(p_i, r, \alpha) = \frac{\widehat{n}(p_i, r, \alpha) - n(p_i, \alpha r)}{\widehat{n}(p_i, \alpha, r)} = 1 - \frac{n(p_i, \alpha r)}{\widehat{n}(p_i, \alpha, r)}. \tag{3.8}$$

Note-se que a r-vizinhança para uma observação p_i contém sempre a observação p_i. Isto, implica que $\widehat{n}(p_i, \alpha, r) > 0$.

Para uma computação mais rápida de MDEF, estima-se $n(p_i, \alpha r)$ e $\widehat{n}(p_i, r, \alpha)$, isto leva à seguinte definição:

Definição 7 *(Contagem e amostragem da vizinhança)*
A contagem da vizinhança (ou αr-vizinhança) é a vizinhança de raio αr, sobre o qual cada $n(p, \alpha r)$ é estimado. A vizinhança da amostra (ou r-vizinhança) é a vizinhança de raio r, sobre o qual separa-se as amostras $n(p, \alpha r)$ para estimar $\widehat{n}(p_i, r, \alpha)$.

Na figura 3.13Representação da n e $\widehat{n}$ - por exemplo $n(p_i, r) = 4, n(p_r, \alpha r) = 1, n(p_1, \alpha r) = 6$ e $\widehat{n}(p_i, r, \alpha) = (1 + 6 + 5 + 1)/4 = 3.25$. [25] figure.caption.39 pode-se observar que o círculo grande limita a r-vizinhança de p_i, enquanto que os círculos menores contêm αr-vizinhanças de vários p's. O principal esquema de deteção de outliers que os autores propõem baseia-se no desvio padrão da contagem da αr-vizinhança sobre a r-vizinhança de p_i, assim tem-se que:

$$\sigma_{\text{MDEF}}(p_i, r, \alpha) = \frac{\sigma_{\widehat{n}}(p_i, r, \alpha)}{\widehat{n}(p_i, r, \alpha)} \tag{3.9}$$

que é o desvio padrão normalizado $\sigma_{\widehat{n}}(p_i, r, \alpha)$ de $n(p_i, \alpha r)$ com $p \in N(p_i, r)$, onde $\sigma_{\widehat{n}}(p_i, r, \alpha) \equiv \sqrt{\frac{\sum_{p \in N(p_i, r)} (n(p, \alpha r) - \widehat{n}(p_i, r, \alpha))^2}{n(p_i, r)}}$.

Usa-se uma vizinhança $0 < \alpha < 1$ para que o cálculo aproximado de MDEF seja rápido. Para reduzir o tempo de execução computacional, determina-se os valores de MDEF e σ_{MDEF}, e explora-se as seguintes propriedades:

- Para cada observação p_i e cada $\alpha, n(p_i, r), \widehat{n}(p_i, r, \alpha)$, o valor de $\text{MDEF}(p_i, r, \alpha)$ e de $\sigma_{\text{MDEF}}(p_i, r, \alpha)$ são constantes por parte de r.
- Em particular, $n(p_i, r)$ e $n(p_i, \alpha r)$ para todos os p's na r-vizinhança de p_i podem apenas mudar quando se aumenta r e faz com que na r-vizinhança de p_i seja adicionado uma nova observação ou na αr-vizinhança de qualquer um dos p's.

Isso, leva à seguinte definição, onde n é o número de observações e $nn(p_i, K)$ é o K-ésimo vizinho mais próximo de p_i.

Definição 8 *(Distância critica)*
Para $1 \leq K \leq n$, chama-se a $d(nn(p_i, K), p_i)$ a distância critica de p_i e $\frac{d(nn(p_i,k),p_i)}{\alpha}$ a distância α-critica de p_i.

Pelas propriedades acima referidas, os autores apenas consideram raios que são críticos ou α-críticos.

Input:

D: base de dados que contém n observações; r: a distância de uma observação; e α: um parâmetro numérico.

Output:

O conjunto de outliers.

O algoritmo LOCI é da seguinte forma:

I. **Pré-processamento**

Para cada $p_i \in D$:

(a) Executa uma pesquisa de um intervalo para $N_i = \{p \in D \mid d(p_i, p) \leq r_{max}\}$;

(b) A partir de N_i, constrói uma lista ordenada L_i das distâncias críticas e α-críticas de p_i

II. **Pós-processamento**

Para cada $p_i \in D$:

(a) Para cada raio $r \in L_i$ (ascendente):

(b) Atualiza $n(p_i, \alpha r)$ e $\widehat{n}(p_i, r, \alpha)$

(c) De n e $\widehat{n}$, calcula o MDEF (p_i, r, α) e $\sigma_{\text{MDEF}}(p_i, r, \alpha)$

(d) Se MDEF $(p_i, r, \alpha) > 3\sigma_{\text{MDEF}}(p_i, r, \alpha)$, então p_i é outlier.

O LOCI conduz a um método aproximado praticamente linear, aLOCI, que proporciona uma deteção rápida e precisa de outliers. O algoritmo aLOCI elimina o alto custo de iterações sobre cada observação na vizinhança de cada p_i. Esta abordagem requer essencialmente apenas contagens em várias escalas, e assim é capaz de superar o problema de multi-granularidade.

Capítulo 4

Aplicação das metodologias a um caso real

Neste capítulo são apresentados os resultados obtidos da aplicação de alguns algoritmos descritos anteriormente. A aplicação dos modelos foi efetuada no software livre *R*, na versão 3.3.3 (6-03-2017) [30]. O erro da máquina associado ao software usado é de 1.110223×10^{-16} (figura 4.1Script usado para determinar o erro da máquina associado ao *R*. figure.caption.40). Todas as funções e bibliotecas usadas serão indicadas no decorrer da análise.

```
> eps<-1.0
> while(eps + 1.0 > 1.0)
+ {
+   eps=eps/2
+ print(eps)
+ }
```

Figura 4.1: Script usado para determinar o erro da máquina associado ao *R*.

Os dados recolhidos e usados como exemplo ao longo desta tese são referentes a uma insígnia do Grupo SONAE (insígnia Y) no âmbito da padaria, pastelaria e cafetaria. A escolha desta insígnia como exemplo na tese deve-se ao facto de ser uma das insígnias em que a tipologia do equipamento é muito abrangente e o número de lojas é considerável. O raciocino apresentado nos próximos capítulos seria o mesmo no caso da aplicação de uma outra insígnia pertencente à *SONAE*.

A base de dados foi exportada de *SAP*, que consiste de um Sistema Integrado de Gestão Empresarial, em formato *EXCEL* no dia 15-02-2017.

4.1 Análise descritiva dos dados

Os dados exportados de *SAP* consistiram de 23275 linhas e 29 colunas, sendo que cada linha da base de dados corresponde a um ativo fixo e as colunas contêm essencialmente informação sobre os ativos fixos. Um extrato da base de dados é apresentado na figura 4.2Extrato da base de dados exportada de *SAP*. figure.caption.41.

Nº inventário	Empr	Cen_Custo	Descritivo Centro de custo	Família	Desc_família	Sub-Família	Desc_sub-família	Localização	Desc_local.	Tipo de Equipamento	Dta invent	Estado do bem
asd	Y	Loja_1	Y Loja_1	M	Sistemas de Informação	M01	POS	1L110		POS Impressora	20.10.2009	Regular
qwe	Y	Loja_1	Y Loja_1	M	Sinalética/Decoração	M01	Reclamos Interiores	1L110	Cafetaria	Reclamo Luminoso - Interior	20.10.2009	Regular
rty	Y	SEDE	Y SEDE	O	Sistemas de Informação	M01	POS	1L110	Cafetaria	POS Monitor	20.10.2009	BOM
uio	Y	SEDE	Y SEDE	M	Utensílios de Frescos	M01	Utens Cafetaria	1L110	Cafetaria	Utensílios - Cafetaria	20.10.2009	BOM
	Y	Loja_5	Y Loja_5	O	Utensílios de Frescos	M01	Utens Cafetaria	1L110	Cafetaria	Utensílios - Cafetaria	20.10.2009	BOM
hhg	Y	Loja_5	Y Loja_5	M	Acrílicos	M01	Acrílicos	1L110	Cafetaria	Acrílicos de Exposição - Charcutari	20.10.2009	Regular
yui	Y	Loja_7	Y Loja_7	M	Utensílios de Frescos	M01	Utens Cafetaria	1L110	Cafetaria	Utensílios - Cafetaria	20.10.2009	Regular
poi	Y	Loja_8	y Loja_8	M	Equipamento de Escritório	M01	Equipamento de Escritório	1L110	Cafetaria	Equipamento de Escritório	20.10.2009	BOM
kkk	Y	Loja_90	y Loja_90	M	Utensílios de Frescos	M01	Inox Cafetaria	1L110		Mesa Inox - Cafetaria	20.10.2009	BOM
	Y	Loja_101	Y Loja_101	M	Mobiliário da Área Social	M01	Mobiliário da Área Social	1L110	Cafetaria	Banco	20.10.2009	BOM
jhj	Y	Loja_95	Y Loja_95	M	Acrílicos de Exposição - Loja	M01	Acrílicos de Exposição - Loj	1L110		Acrílicos de Exposição - Loja	20.10.2009	BOM
frt	Y	Loja_95	Y Loja_95	T	Equip Inox	M01	Mesa Inox - Cafetaria	1L110	Cafetaria	Mesa Inox - Cafetaria	20.10.2009	BOM
iop	Y	Loja_13	Y Loja_13	M	Mobiliário de Escritório	M01	Mobiliário de Escritório	1L110	Cafetaria	Equipamento de Escritório	20.10.2009	BOM
hjh	Y	Loja_2	Y Loja_2	H	Mobiliário da Área Social	H08	Mobiliário da Área Social	1L110		Banco	20.10.2009	BOM
hhg	Y	Loja_2	Y Loja_2	H	Mobiliário de Escritório	H08	Mobiliário de Escritório	1L110	Cafetaria	Cadeira	20.10.2009	BOM
ytr	Y	Loja_16	Y Loja_16	T	EQUIPAMENTOS DE APOIO Á OPERAÇ	H08	Equipamento de hotelaria	1L110	Cafetaria	Abre Latas	20.10.2009	BOM
	Y	SEDE	Y SEDE	H	Carpintarias	H08	Banco	1L110	Armazem	Banco	20.10.2009	BOM
...	...	...	...	...	...	...	...	...	...	...	...	...

Sbnº	Dt Aq.Ori.	Ini.Dpr.no	Qtd	UBM	Fornecedor	Val.aquis.a	Moeda	Val.cont.fi.exer	Moeda1	Det.ctas	Código DGCI	Observações	Fabricante do imobilizado	AA-Benefícios	Data movimentação Permitida
0	31.05.2009	01.05.2009	1		Fornecedor_1	660,89	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_3	66,89	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_3	1000,3	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_4	10,5	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_5	10,5	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_6	90	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_70	66,89	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_8	20,5	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_9	100,5	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_10	52,25	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_15	300,25	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_15	40,2	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_13	66,89	EUR	0	EUR	43504351	2430				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_14	15,08	EUR	0	EUR	43304331	1655				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_5	140,48	EUR	0	EUR	43304331	1655				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_5	15,23	EUR	0	EUR	43304331	1655				00.00.0000
0	31.05.2009	01.05.2009	1		Fornecedor_1	140,48	EUR	0	EUR	43304331	1655				00.00.0000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figura 4.2: Extrato da base de dados exportada de *SAP*.

Na tabela B.1Descrição das variáveis .table.caption.85, em anexo BDescrição das variáveis do conjunto de dadosAnexo.a.B, apresenta-se uma descrição do conjunto de variáveis levantadas. O objetivo do estudo consiste em construir um sistema de controlos e de alertas sempre que se verifiquem desvios significativos no que se refere a inventários de ativos fixos, isto é pretendem-se identificar as lojas cujo número de equipamentos, por metro quadrado, é atípico. As lojas que são destacadas como sendo atípicas, são lojas que têm um número de equipamentos muito elevado/reduzido, por metro quadrado, e têm de ser devidamente analisadas pela equipa da "Gestão de Ativos Fixos".

Os modelos focaram-se nos equipamentos etiquetáveis contidos nos equipamentos básicos, isto é máquinas, ferramentas, equipamentos de decoração e outros bens com os quais se realiza a extração, transformação e elaboração dos produtos ou a prestação dos serviços. A equipa da Inventariação, que tem como uma das funções fazer auditorias às lojas, quando visita as lojas apenas faz o levantamento dos equipamentos etiquetáveis. Pela base de dados da figura 4.2Extrato da base de dados exportada de *SAP*. figure.caption.41, pode-se observar que na

insígnia em estudo encontram-se equipamentos que não são etiquetáveis, tais como tabuleiros lisos e utensílios. É de salientar também que na variável Descritivo do Centro de Custo pode-se encontrar lojas e sedes, mas para a presente análise só as lojas têm interesse. Antes de construir o modelo foi feito um tratamento de dados de forma a eliminar os equipamentos não etiquetáveis e na variável Descritivo do Centro de Custo apenas permanecer as lojas, reduzindo o número de linhas. Esse tratamento foi executado no programa *EXCEL*, em que se eliminaram também as variáveis que não vão ser utilizadas no modelo. Para tal criou-se uma aplicação em *EXCEL*, figura 4.3Aplicação para o tratamento de dados em *EXCEL*. figure.caption.42, com o auxílio de *Macros* e código *VBA* (Visual Basic for Applications).

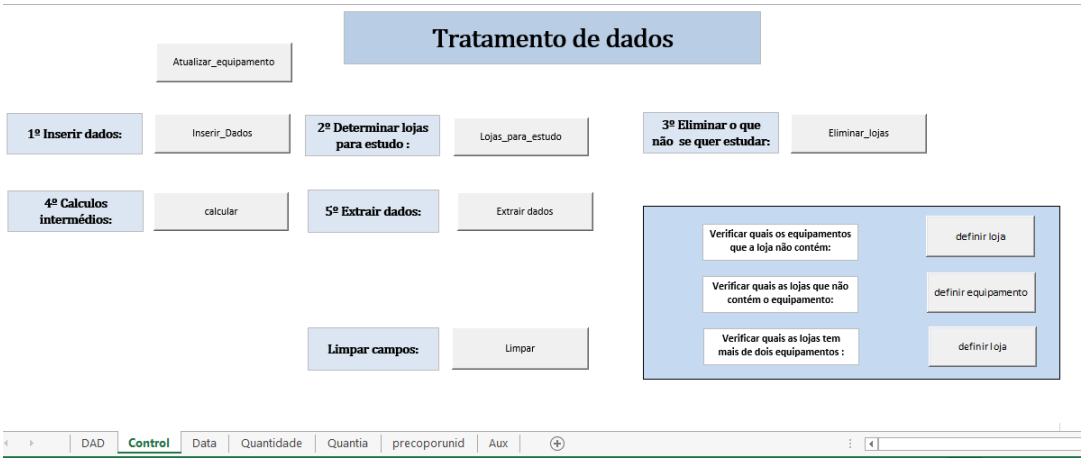

Figura 4.3: Aplicação para o tratamento de dados em *EXCEL*.

Este ficheiro *EXCEL* contém sete folhas:

- **DAD**: referente à base de dados exportada diretamente pela query do programa *SAP*, uma vez que as *Macros* estão programadas para esse formato;
- **Control**: como o nome indica é nesta folha que se controlam as restantes folhas através dos botões, estes devem ser clicados por uma certa ordem;
- **Data**: enquanto no início não se encontra preenchida, no decorrer do tratamento de dados, obtêm-se as variáveis de maior importância para a análise;
- **Quantidade**: inicialmente estará em branco, mas no decorrer do tratamento é feita a identificação do número de cada tipo de equipamento existente em cada loja, que será extraída para posteriormente ser utilizada no software *R*;
- **Quantia**: é semelhante à anterior, mas neste caso é referente ao valor do equipamento (de acordo com as variáveis Val. aquis. atual e tipo de equipamento);

- **Precoporunid**: inicialmente encontra-se em branco, mas no desenrolar do programa será preenchido o preço unitário de cada tipo de equipamento, por loja; e

- **Aux**: é uma folha auxiliar, que serve de suporte para a programação das restantes folhas.

Na folha Control os botões (figura 4.3Aplicação para o tratamento de dados em *EXCEL*. figure.caption.42) têm de ser pressionados pela seguinte ordem:

1. Antes de começar a tratar uma nova insígnia, clica-se no botão **Limpar**, que elimina os campos preenchidos pelo programa da última vez que se utilizou a aplicação;

2. Em todas as análises devem-se atualizar os equipamentos, uma vez que estão sempre a surgir equipamentos novos, para tal seleciona-se o botão **Atualizar_equipamento**;

3. Clica-se no botão **Inserir_Dados**, a *Macro* acrescenta na folha DAD duas colunas, uma referente à divisão da loja e outra que classifica os equipamentos se são etiquetáveis ou não. A *Macro* com o recurso da folha auxiliar elimina os equipamentos não etiquetáveis, uma vez que o estudo só se foca nos equipamentos etiquetáveis. De seguida, a folha Data fica preenchida com as seguintes variáveis: Cent_Custo, localização, Descritivo do Centro de Custo, Tipo de Equipamento (apenas aparecem os equipamentos etiquetáveis), Qtd e Val. aquis. atual;

4. Seleciona-se o botão **Lojas_para_estudo**, que acrescenta uma coluna onde o utilizador seleciona apenas as lojas que pretende analisar;

5. Clica-se no botão **Eliminar_lojas**, para eliminar as lojas cujo utilizador selecionou;

6. Pressiona-se no botão **calcular**, que preenche as seguintes folhas: quantia, quantidade e precoporunidade. Nesta fase obtêm-se os nomes das lojas (observações) e os nomes dos tipos de equipamentos (variáveis);

7. Seleciona-se o botão **Extrair dados**, que tem como função extrair dois ficheiros para uma pasta, quantidade e quantia, que serão posteriormente lidos no programa *R*.

Como se pode observar, na folha Control existem três botões que permitem uma análise mais detalhada das lojas, os nomes dos equipamentos duplicados e os nomes dos equipamentos em falta numa determinada loja, e os nomes das lojas que não contêm um determinado equipamento.

Nas folhas quantidade, quantia e precoporunid foram tidas em conta as variáveis Cen_Custo (nome das lojas), o tipo de equipamento (nome dos equipamentos etiquetáveis), a quantidade

dos bens e a respetiva quantia. De forma a simplificar a explicação do que foi feito nessas folhas, observa-se o seguinte exemplo:

	Mesa			Bancada			Cadeira		
	Quantidade	Quantia (€)	Precoporunid (€)	Quantidade	Quantia (€)	Precoporunid (€)	Quantidade	Quantia (€)	Precoporunid (€)
Loja_1	3	400	133,33	3	1000	333,33	3	10	3,33
Loja_2	5	500	100	2	500	250	5	35	7
Loja_3	7	800	114,29	6	900	150	9	20	2,22
valor médio	5		115,87	3,67		244,44	5,67		4,18
Quantia média por tipo de equipamento	579,35 €			897,10 €			23,70 €		

Figura 4.4: Exemplo do procedimento para preencher as folhas quantidade, quantia e precoporunid, no *EXCEL*.

Inicialmente, foi feita a identificação do número de cada tipo de equipamento existente em cada loja e, no fim, calculou-se a média da quantidade de cada tipo de equipamento numa insígnia Y. De seguida, fez-se a soma da quantia de cada tipo de equipamento, por loja. De forma a obter-se o valor unitário de cada equipamento, por loja, determinou-se o quociente entre a quantia e a quantidade, e calculou-se a média do valor unitário de cada equipamento, obtendo-se assim o valor médio desse equipamento na insígnia em estudo. No final, calculou-se o produto entre a média da quantidade do equipamento e a média do valor unitário desse equipamento, que resulta no valor médio desse equipamento numa insígnia Y. Este valor será utilizado nas abordagens para criar o modelo alarmístico, com o intuito de diminuir a tipologia do equipamento por forma a manter apenas os equipamentos que têm um número elevado ou um valor da aquisição elevado, numa insígnia Y.

Para simplificar, apresentamos, por exemplo, o que se fez no caso das **cadeiras**:

1. Somou-se a quantidade de cadeiras, nas lojas 1, 2 e 3;

2. De seguida, determinou-se a média das cadeiras na insígnia Y, $\frac{3+5+9}{3} = 5$, isto é, em média numa insígnia Y têm-se 5 cadeiras;

3. Somou-se a quantia desse equipamento, nas lojas 1, 2 e 3;

4. Determinou-se o valor unitário médio, isto é, calculou-se o quociente entre a quantia e a quantidade das cadeiras, por exemplo na Loja_1, $\frac{10}{3} = 3.33$ €;

5. De seguida, determinou-se a média do valor unitário das cadeiras, isto é, $\frac{3.33+7+2.22}{3} = 4.18$. Este valor indica que, em média, numa insígnia Y, uma cadeira custou 4.18 €; e

6. Por fim, fez-se o produto entre a média da quantidade e a média do valor unitário, isto é, $5 \times 4.18 = 23.70$, este valor declara que, em média, cada loja da insígnia em estudo tem 23.70 € em cadeiras.

Para a construção do modelo, utilizaram-se três abordagens implementadas no software *R*, considerando-se a base de dados quantidade e a base de dados quantia, e que foram denominadas, respetivamente, por Base_Total e Data_Total. Estas bases de dados, após o tratamento, contêm 108 observações referentes às lojas pertencentes à insígnia Y e 164 variáveis referentes aos tipos de equipamentos. De forma a melhorar o modelo inclui-se na análise uma outra base de dados que contém a informação da área de cada loja e o tipo de intervenção feita nos últimos dois anos.

Note-se que as lojas que não contêm a informação da área são lojas que já encerraram e, como tal, não são importantes para a análise, ou lojas que abriram há pouco tempo e não existem todos os elementos necessários para análise. Eliminaram-se as lojas que abriram recentemente (Loja_70, Loja_82 e Loja_108) devido a faltar algumas informações. A Loja_104 também foi eliminada da base de dados, porque é uma loja que já encerrou e não tem importância para análise. Estas quatro lojas foram eliminadas das bases de dados.

As bases de dados, Base_Total e Data_Total, para os seguintes capítulos têm a seguinte dimensão: 104 observações e 164 variáveis. Para uma melhor compreensão dos dados, observa-se como as 4 primeiras variáveis se encontram distribuídas através do comando *boxplot()* em *R*:

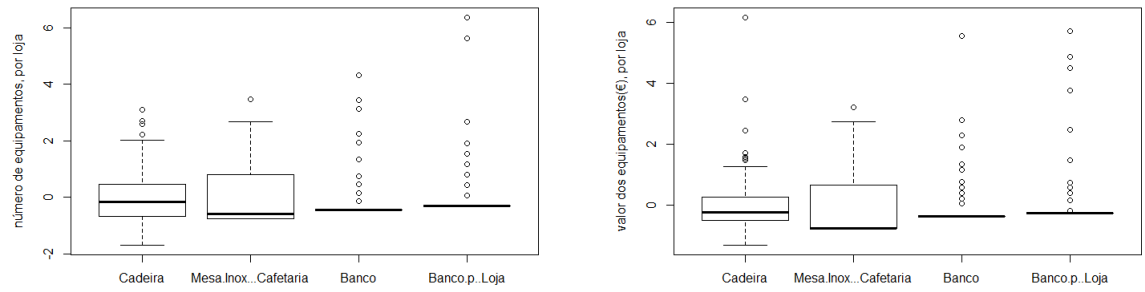

Figura 4.5: Boxplots referentes às primeiras quatro variáveis, normalizadas, da Base_Total e da Data_Total, respetivamente.

4.1.1 Abordagem 1

Numa primeira abordagem, para a deteção de outliers, agruparam-se as 164 variáveis pela sua tipologia, isto é agruparam-se os equipamentos de acordo com as suas funções numa insígnia Y, ficando-se assim com 9 variáveis: **Decoração**, onde agruparam-se os equipamentos de decoração, tais como mesas, cadeiras, quadros, entre outros; **Extra**, encontram-se os equipamentos não tão necessários para o funcionamento de uma insígnia Y; **Frio**, agruparam-se os

equipamentos de frio, tais como frigoríficos, vitrinas, entre outros ; **Quente**, compilaram-se os equipamentos de quente, tais como fogões, fornos, entre outros; **Bancada**, aglomeraram-se os equipamentos de arrumação, tais como bancadas, estantes, entre outros; **Hardware**, como o nome indica agruparam-se todos os hardwares e **Peq_equipamentos**, juntaram-se os equipamentos pequenos, tais como caixotes do lixo, escadotes, barreiras, entre outros.

A aplicação desta abordagem foi feita à Base_Total e à Data_Total, obtendo-se assim a Base e a Data, respetivamente, cuja dimensão é de 104 observações e 9 variáveis. A estas novas bases de dados foram-lhes aplicados alguns dos algoritmos de deteção de outliers, estudados nos capítulos anteriores.

As técnicas de deteção de outliers, implementas nas bases de dados, que se seguem são apenas técnicas não supervisionadas, pois fazem a suposição implícita de que as observações normais são muito mais frequentes do que os outliers.

Abordagem 1 referente à base de dados Base

Na base de dados **Base**, o objetivo é analisar o número de equipamentos nas lojas.

Na aplicação da técnica de deteção de outliers baseada em métodos estatísticos não paramétricos destaca-se o boxplot. O boxplot permite avaliar a simetria dos dados, a dispersão e a existência de outliers. Através do comando *boxplot( )* obtém-se:

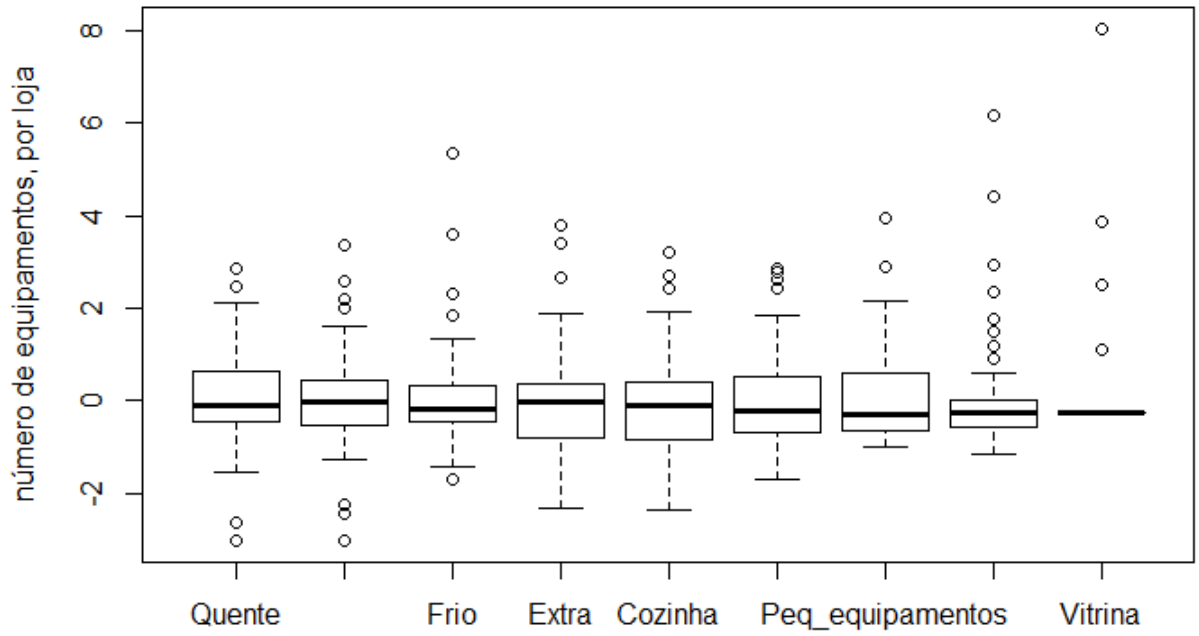

Figura 4.6: Boxplots referente às variáveis, normalizadas, da base de dados Base.

Pela figura 4.6Boxplots referente às variáveis, normalizadas, da base de dados Base. figure.caption.46 observa-se que existem muitos outliers em todas as variáveis, o que não ajuda a concluir nada sobre os dados.

De seguida, aplicaram-se técnicas de deteção de outliers baseadas em clustering, mais concretamente, o algoritmo FindCBLOF. Este algoritmo, que primeiro particiona o conjunto de dados em clusters através do algoritmo K-médias, de seguida, para cada ponto da base de dados, o valor de CBLOF é calculado. As observações em que o valor CBLOF seja elevado consideram-se outliers. De forma a implementar o K-médias estudaram-se as distribuições das 9 variáveis. O melhor valor de K a aplicar à base de dados foi $K = 3$, ou seja, particionaram-se as observações em três grupos. Para tal, executou-se em R o comando *kmeans(Base,3)*, recorrendo à biblioteca *stats*, tendo-se obtido o seguinte particionamento das lojas:

Cluster	**1**	**2**	**3**
Nº de lojas	10	57	37

Tabela 4.1: Particionamento das 104 lojas, aplicando o algoritmo K-médias.

Posteriormente, aplicou-se o comando, em anexo CAlgoritmo FindCBLOFAnexo.a.C, para determinar o algoritmo FindCBLOF, uma vez que o *R* não contém em nenhuma livraria o algoritmo FindCBLOF.

As cinco lojas cujo valor do CBLOF, determinado pelo algoritmo FindCBLOF, é maior: "Loja_29", "Loja_62", "Loja_71", "Loja_4" e "Loja_15". O valor do CBLOF foi calculado atribuindo a cada observação um fator de outlier, que é medido pelo tamanho do cluster ao qual a observação pertence, e a distância entre as observações e o cluster mais próximo. A distância entre as observações e o cluster mais próximo foi determinada pela variável área, de cada loja.

Por fim aplicaram-se as técnicas de deteção de outliers baseadas na proximidade, sendo estas classificadas em distância e em densidade. A técnica baseada na distância mais aplicada é o K vizinho mais próximo, KNN, que determina o score de outlier de uma observação, com base nas distâncias dos K vizinhos mais próximos.

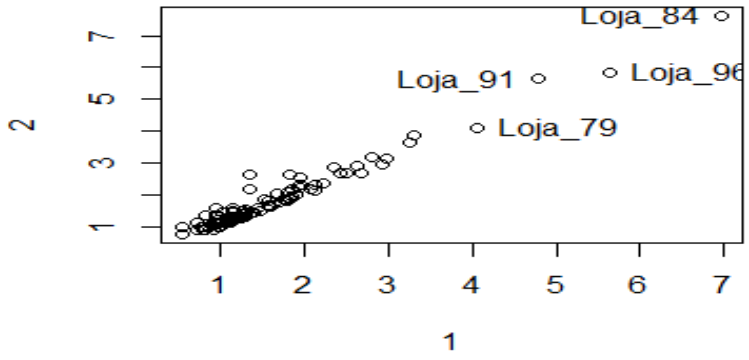

Figura 4.7: Aplicação do método KNN à base de dados, Base.

Para efetuar esta tarefa foi essencial a função *kNNdist( )* da livraria *dbscan*. Aplicou-se ao método KNN o parâmetro $K = 2$, para obter-se o gráfico 4.7Aplicação do método KNN à base de dados, Base. figure.caption.48. Foram também testados diferentes valores de K mas obteve-se sempre as mesmas conclusões. Pelo método KNN determinaram-se as seguintes lojas como possíveis outliers: "Loja_84", "Loja_96", "Loja_91" e "Loja_79". As lojas destacadas têm a distância, aos seus 2-vizinhos mais próximos, maior do que as restantes lojas. Neste método não se cruzou a informação com a área, apenas obtiveram-se os outliers devido ao número de equipamentos ser muito elevado ou reduzido.

De seguida, aplicaram-se aos dados técnicas de deteção de outliers baseadas na densidade. As técnicas aplicadas são: o LOF e o LOCI. Nestas técnicas, uma observação 'normal' (não

outlier) tem a sua densidade local semelhante a dos seus vizinhos, enquanto que uma observação outlier tem a sua densidade local menor que a dos seus vizinhos.

A primeira técnica aplicada foi o LOF. O valor desta técnica é maior num outlier de que numa observação 'normal'. Para determinar o valor de LOF em *R*, recorreu-se ao comando *lof( )* da livraria *Rlof*. O valor de LOF foi determinado, para cada observação, aplicando diferentes valores de K (3, 6, 9, 12, 15, 20, 25, 30, 35 e 40), isto é, aplicaram-se diferentes 'MinPts' de modo a interpretar o comportamento dos dados à medida que se aumentava o 'MinPts'. Representou-se o valor de LOF de cada observação para cada 'MinPts', num gráfico, para tal foi necessário instalar a livraria *ggplot2*.

Pela aplicação do comando anterior, obteve-se o seguinte gráfico:

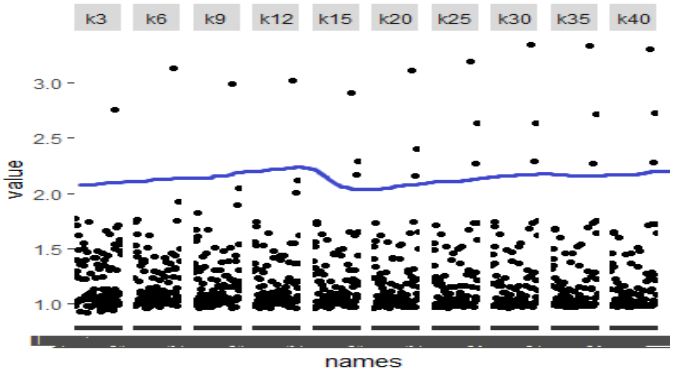

Figura 4.8: Aplicação do método LOF à base de dados, Base.

Através do gráfico 4.8Aplicação do método LOF à base de dados, Base. figure.caption.49, constata-se que para o valor de $K \geq 25$ o valor de LOF de cada observação tende a estabilizar. Os outliers são as observações cujo valor de LOF é maior, e como se verifica pelo gráfico, não é simples determinar um valor de corte para os obter. A melhor forma para determinar os outliers foi traçar uma linha (gráfico 4.8Aplicação do método LOF à base de dados, Base. figure.caption.49) de forma a interpretar as lojas cujo valor está acima dessa linha. Para $K = 3$, $K = 6$, $K = 9$ e $K = 12$ obteve-se a "Loja_84", enquanto que para $K = 15$ até ao $K = 40$ obteve-se a "Loja_84", a "Loja_96" e a "Loja_91". A partir do $K = 15$ pode-se verificar pelo gráfico que o valor de LOF forma 'três grupos', dessa forma consideram-se as três observações cujo valor de LOF é, notoriamente, maior do que os restantes. Conclui-se assim, que os possíveis outliers são: "Loja_84", "Loja_96" e "Loja_91".

Por fim, aplicou-se a técnica LOCI. Esta técnica também é baseada na densidade mas, contrariamente ao LOF, não é necessário definir um valor de corte, o próprio algoritmo determina quais as observações que são consideradas outliers. Na aplicação do LOCI recorreu-se à função *LOCI()* e à *prabclust(prabinit( ))* das livrarias *SMLoutliers* e *prabclus*, respetivamente,

obtendo-se as lojas outliers e o gráfico seguinte.

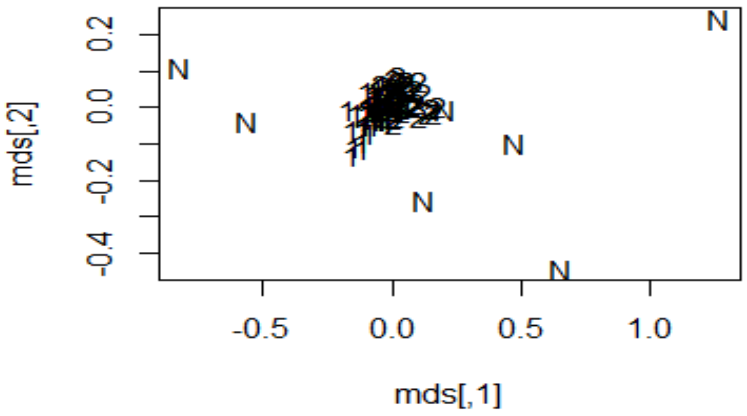

Figura 4.9: Aplicação do método LOCI à base de dados, Base.

Salienta-se ainda que, para se obter o gráfico, é utilizada a distância de Jaccard. Neste gráfico, os outliers são representados por N e as restantes observações são particionadas, e representadas pelos clusters 1 e 2.

As lojas que o valor LOCI retornou como sendo possíveis outliers são: "Loja_79", "Loja_84", "Loja_89", "Loja_91", "Loja_96", "Loja_97" e "Loja_98". Note-se que foi considerado o parâmetro $alpha = 0.5$, tendo sido testados outros valores, mas à medida que o parâmetro era aumentado, mais lojas eram retornadas como outliers, quando o parâmetro era diminuído, menos lojas retornavam como possíveis outliers. Por este motivo, considerou-se $alpha = 0.5$, uma vez que este valor, devolvia um número considerável de lojas como possíveis outliers.

Na tabela D.1Lojas alarmísticas obtidas na base de dados Base .table.caption.86, em anexo DTabelas da abordagem 1Anexo.a.D, apresenta-se de uma forma resumida as lojas outliers, para cada técnica, da base de dados, Base.

Numa primeira abordagem da base de dados, Base, conclui-se que a técnica FindCBLOF não está em conformidade com as demais. As restantes técnicas retornam lojas outliers em comum, dessas destacam-se "Loja_84", "Loja_96", "Loja_91" e Loja_79".

Abordagem 1 referente à base de dados Data

Na base de dados **Data**, estuda-se, em cada loja, o valor dos equipamentos.

Nesta abordagem apenas não se aplicará o algoritmo FindCBLOF, uma vez que se teria de

particionar a base de dados conforme o valor dos equipamentos. Aos métodos aplicados não se cruzou a informação da área das lojas, tendo-se apenas estudado o valor dos equipamentos em cada loja. Através do comando *boxplot( )* que permite avaliar a simetria dos dados, a dispersão e a existência de outliers obtém-se a seguinte figura:

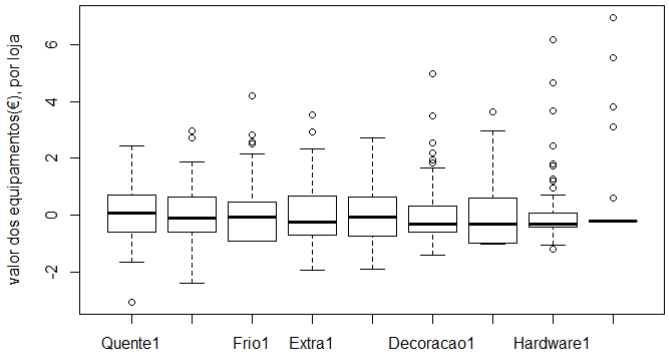

Figura 4.10: Boxplots referentes às variáveis, nrmalizadas, da base de dados Data.

À semelhança dos boxplots obtidos para a Base, observa-se que nestes boxplots também existem muitos outliers em todas as variáveis, o que não ajuda a concluir nada sobre os dados.

De seguida aplicaram-se as técnicas de deteção de outliers baseadas na proximidade que são classificadas em distância e densidade. A técnica baseada em distância mais aplicada é K vizinhos mais próximos, KNN.

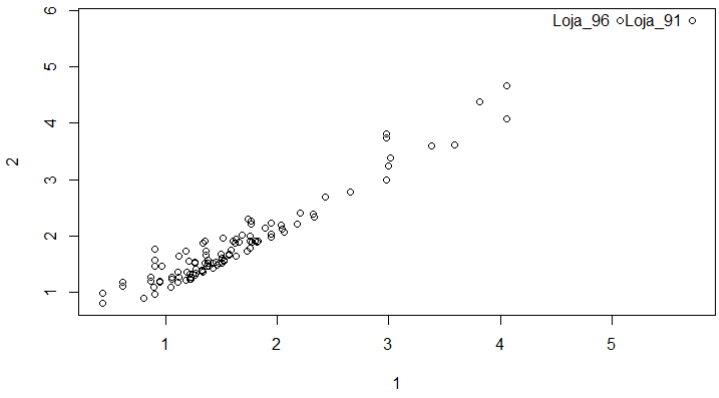

Figura 4.11: Aplicação do método KNN à base de dados, Data.

Pelo método KNN, determinou-se as lojas "Loja_96" e "Loja_91", como sendo as possíveis lojas outliers, lojas estas cuja distância aos seus K-vizinhos mais próximos é maior, comparativamente com as demais. Aplicou-se ao método KNN o parâmetro $K = 2$, para obter-se o

gráfico 4.11Aplicação do método KNN à base de dados, Data. figure.caption.53, também foram testados diferentes valores de K mas obteve-se sempre as mesmas conclusões.

De seguida, aplicou-se aos dados técnicas de deteção de outliers baseados na densidade, tais técnicas são o LOF e o LOCI. A primeira técnica aplicada foi o LOF. O valor de LOF à semelhança com a análise anterior, foi determinado para cada observação, aplicando diferentes K's. Representou-se o valor de LOF de cada observação para cada K, no seguinte gráfico:

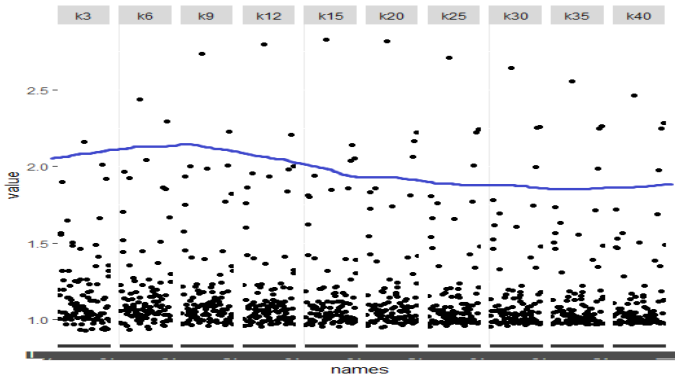

Figura 4.12: Aplicação do método LOF à base de dados, Data.

Através do gráfico 4.12Aplicação do método LOF à base de dados, Data. figure.caption.54, pode-se notar que para o valor de $K \geq 25$ o valor de LOF para cada observação tende a estabilizar. À semelhança da análise anterior, a melhor forma para determinar os outliers foi traçar uma linha (gráfico 4.12Aplicação do método LOF à base de dados, Data. figure.caption.54) de forma a interpretar as lojas cujo valor está acima dessa linha. Para $K = 3$ obteve-se apenas a "Loja_50" como possível outlier. Já para $K = 6$, $K = 9$ e $K = 12$ têm-se a "Loja_40" e a "Loja_91" como possíveis outliers, enquanto que para $K = 15$ até $K = 40$ obteve-se a "Loja_40", a "Loja_91", a "Loja_96" e a "Loja_88". Conclui-se assim que as possíveis lojas outliers, no algoritmo LOF, são: "Loja_50", "Loja_40", "Loja_91", "Loja_96" e "Loja_88".

Por fim, aplicou-se a técnica LOCI e obtive-se o seguinte gráfico:

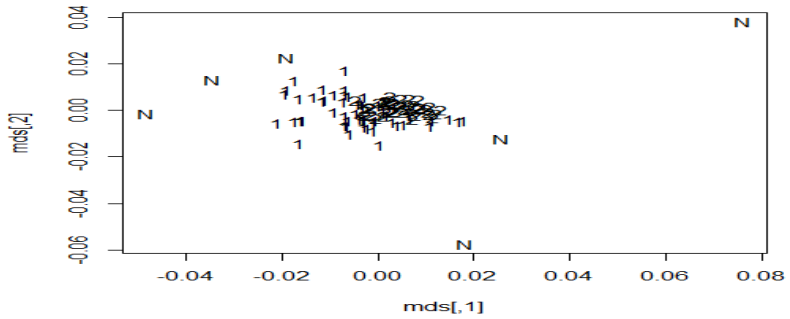

Figura 4.13: Aplicação do método LOCI à base de dados, Data.

As lojas que a técnica LOCI retornou como sendo possíveis outliers são: "Loja_79", "Loja_84", "Loja_91", "Loja_96", "Loja_105" e "Loja_106". Note-se que se considerou o parâmetro $alpha = 0.4$, onde retornou um número considerável de outliers,

Na tabela D.2Lojas alarmísticas obtidas na base de dados Data .table.caption.87, em anexo DTabelas da abordagem 1table.caption.86, apresenta-se de uma forma resumida as lojas outliers, para cada técnica, da base de dados, Data. Na base de dados Data, destaca-se a Loja_91" e a Loja_96".

Conclui-se que, tanto na base de dados Base como na Data, as lojas que se destacam em ambas as análises são: "Loja_84", "Loja_91", "Loja_96" e "Loja_79".

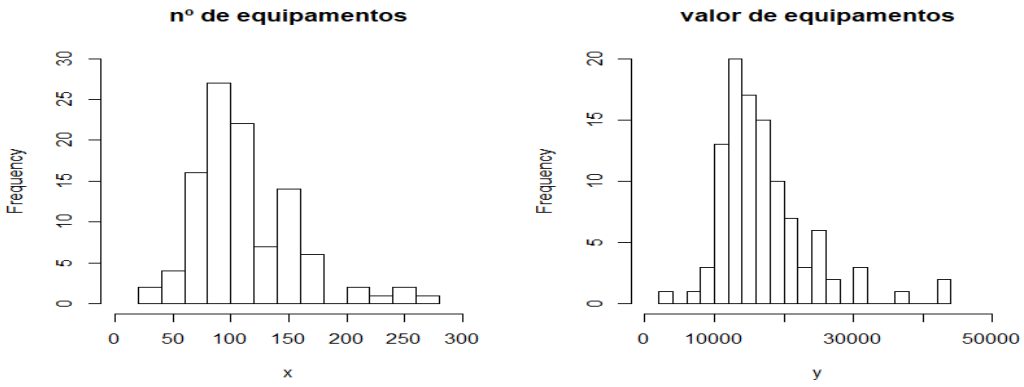

Figura 4.14: Histogramas referentes ao número de equipamentos e ao valor dos equipamentos, nas lojas, respetivamente.

Observa-se pelo histograma referente ao número de equipamentos, que as lojas com mais de 200 equipamentos, por ordem crescente, são: "Loja_89", "Loja_98", "Loja_97", "Loja_79", "Loja_91" e "Loja_96". No histograma referente ao valor dos equipamentos, as lojas com mais de 30000 € em equipamentos são: "Loja_89", "Loja_98", "Loja_79", "Loja_106", "Loja_96" e

"Loja_91".

Conclui-se que das lojas destacadas na tabela D.3Lojas alarmísticas obtidas na abordagem 1. table.caption.88, em anexo DTabelas da abordagem 1table.caption.87, e pela figura 4.14Histogramas referentes ao número de equipamentos e ao valor dos equipamentos, nas lojas, respetivamente. figure.caption.56 as mais alarmísticas são: "Loja_91" e "Loja_96". Ambas as lojas destacadas sofreram remodelação, o que pode justificar o facto de ter muitos equipamentos, por ainda não terem feito as movimentações devidas de alguns equipamentos (vendas e abates).

4.1.2 Abordagem 2

Pela abordagem anterior obtiveram-se possíveis lojas outliers, mas não se cruzou os dados com a área das mesmas, apenas se estudou o número de equipamentos e o seu valor, em cada loja. Salienta-se que o objetivo do estudo consiste em determinar as lojas mais alarmísticas, isto é, as lojas cujo número de equipamentos, por metro quadrado, é muito reduzido ou muito elevado. Para tal, aplicou-se uma segunda abordagem. Nesta abordagem, o objetivo consiste em agrupar as lojas, pelo número de equipamentos, e depois determinar as lojas atípicas dentro de cada grupo.

A Base_Total e a Data_Total contêm 104 observações e 164 variáveis. A forma mais correta será reduzir o número de variáveis para que a aplicação de qualquer algoritmo seja mais fiável, uma vez que se aplicará métodos de clustering, para particionar as observações. Para tal, observa-se o exemplo de figura 4.4Exemplo do procedimento para preencher as folhas quantidade, quantia e precoporunid, no *EXCEL*. figure.caption.43, em que se determinou no tratamento de dados o parâmetro da quantia média por tipo de equipamento, com o objetivo de ser usado para corte das variáveis, isto é, equipamentos cuja quantia média na insígnia Y seja inferior a 30 € eliminam-se da Base_Total e da Data_Total. O valor de corte foi de 30 € por forma a garantir que ficavam para análise os equipamentos de grande quantidade ou muito caros, em cada insígnia Y.

As novas bases de dados, Base_1 e Data_1, contêm 104 observações (lojas pertencentes à insígnia Y) e 40 variáveis (tipo de equipamentos cujo valor médio é superior a 30 €). Para uma melhor compreensão dos dados, observa-se como as 4 primeiras variáveis se encontram distribuídas, em cada base de dados, através do comando *boxplot()* em R:

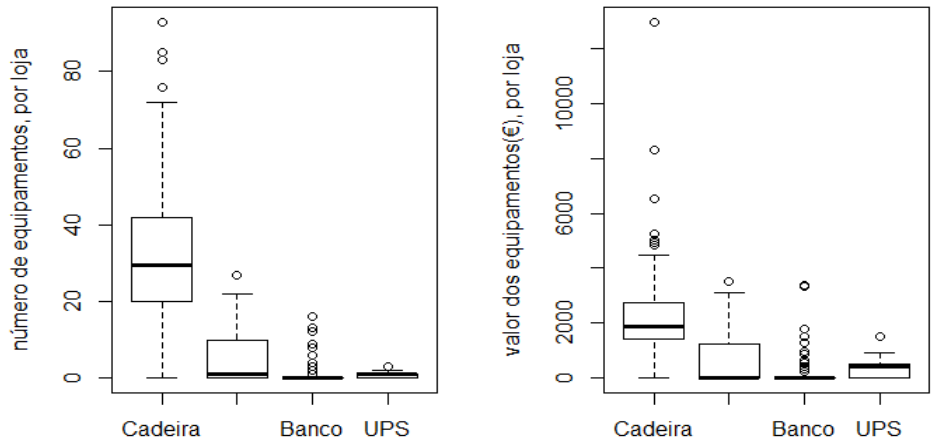

Figura 4.15: Boxplots referentes às quatro primeiras váriaveis da Base_1 e da Data_1, respetivamente.

De seguida, estudou-se o melhor algoritmo de clustering para aplicar aos dados. Um bom algoritmo de agrupamento forma grupos cuja semelhança entre elementos no mesmo grupo é elevada e a semelhança entre elementos de grupos diferentes é reduzida. De forma a determinar o melhor valor de K (número de grupos a formar) para os métodos foi essencial a aplicação do método do cotovelo. Para aplicar o método do cotovelo foi necessário utilizar a função *fviz_nbclust( )* da livraria *factoextra*, para se obterem os seguintes gráficos.

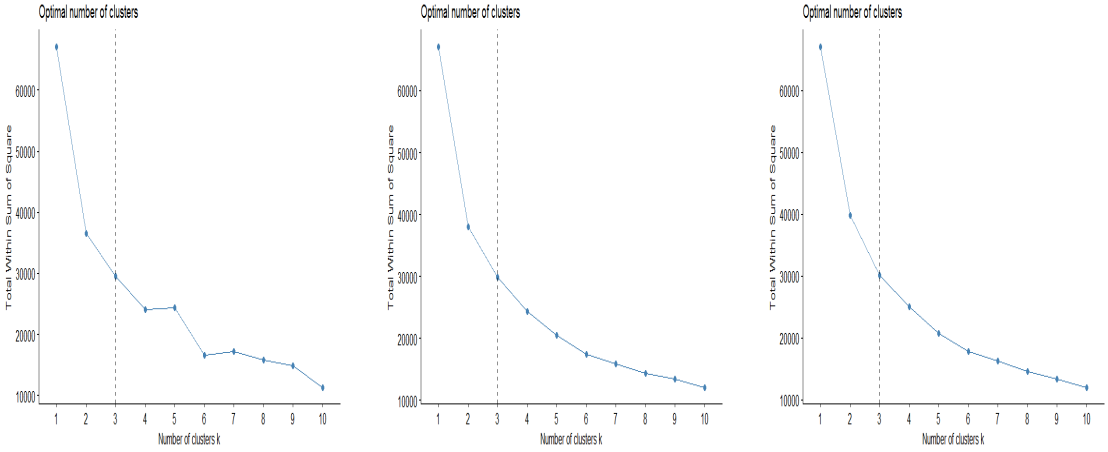

Figura 4.16: Determinação de K nos métodos K-médias, PAM e hierárquico aglomerativo (da esquerda para a direita).

No gráfico procura-se o "cotovelo" para definir qual é o número aceitável de K (clusters) a a formar, com base nos dados da amostra. Este método aumenta a quantidade de clusters a partir

de 1 e analisa o resultado a cada incremento. Pelos gráficos acima, conclui-se que $K = 3$ será o melhor valor de K a aplicar nos algoritmos, verificou-se também as distribuições de alguns equipamentos, para tal, recorreu-se ao histograma através do comando *hist()* e obteve-se as seguintes representações:

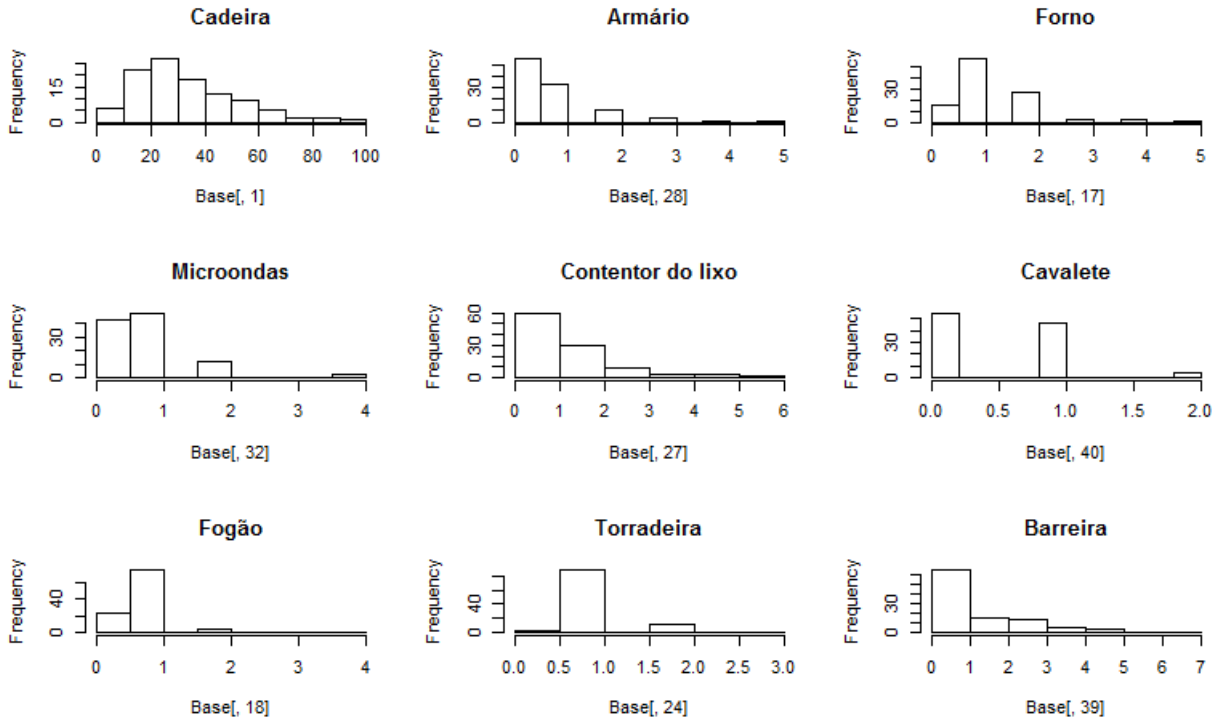

Figura 4.17: Representações dos histogramas de alguns equipamentos.

Pela figura 4.17Representações dos histogramas de alguns equipamentos. figure.caption.59 pode-se concluir que muitas lojas em relação ao número de equipamentos ou têm um, ou têm em duplicado ou não têm o equipamento, constatando-se que dividir as observações em três grupos será o ideal. Aplicou-se também a análise o parâmetro $K = 4$ de forma a comparar os resultados dos métodos.

No método agrupamento hierárquico aglomerativo utilizou-se a distância *Euclideana* e aplicaram-se as seguintes ligações: average, single, ward.D, complete e centroid. No fim, determinou-se o coeficiente de correlação cofenética (CCC) para concluir o melhor método. Um valor alto para CCC é considerado como uma medida de classificação bem-sucedida.

Determinou-se o método hierárquico aglomerativo, recorrendo à função *hclust( )* da livraria *stats*, tanto para $K = 3$ como para $K = 4$, obtendo-se os gráficos seguintes:

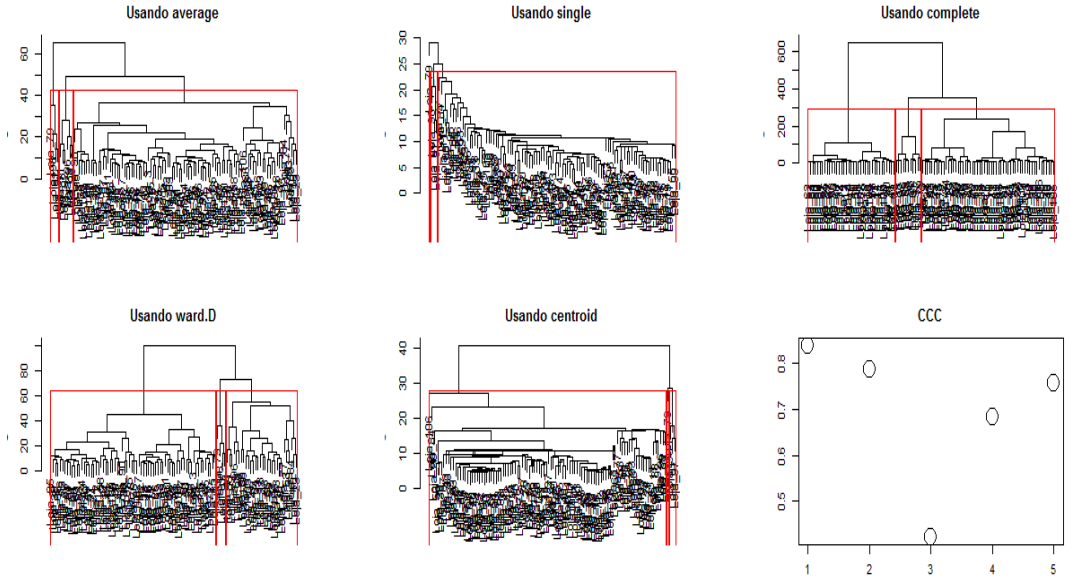

Figura 4.18: Representação do método hierárquico aglomerativo com $K = 3$.

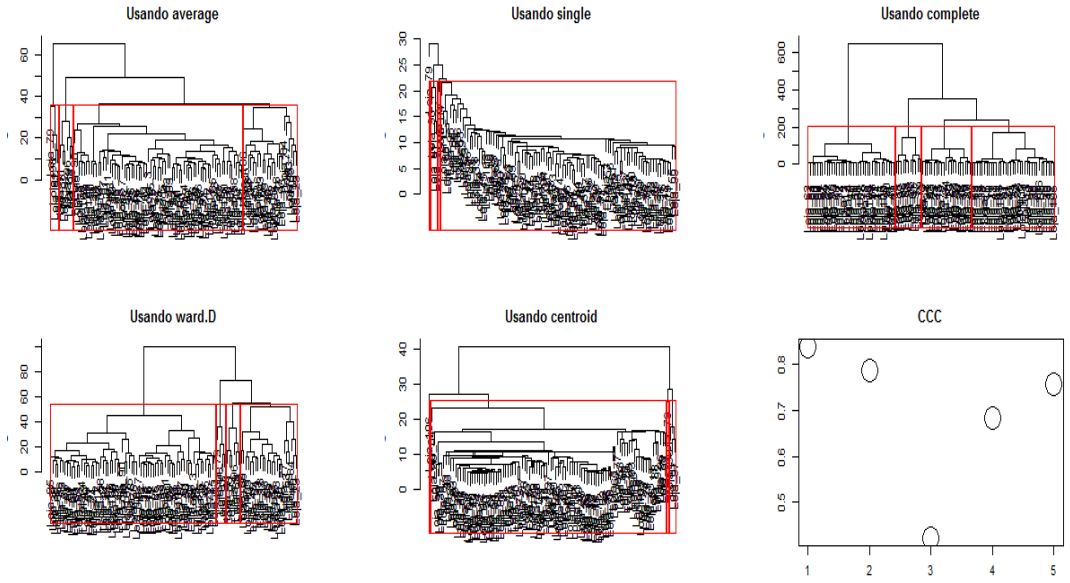

Figura 4.19: Representação do método hierárquico aglomerativo com $K = 4$.

Pode-se observar que, na figura 4.18Representação do método hierárquico aglomerativo com $K = 3$. figure.caption.60 e na 4.19Representação do método hierárquico aglomerativo com $K = 4$. figure.caption.61, nenhuma ligação agrupa as observações de forma homogénea. Nas duas figuras a melhor ligação é a average, uma vez que o valor do coeficiente de correlação cofenética é superior as restantes ligações.

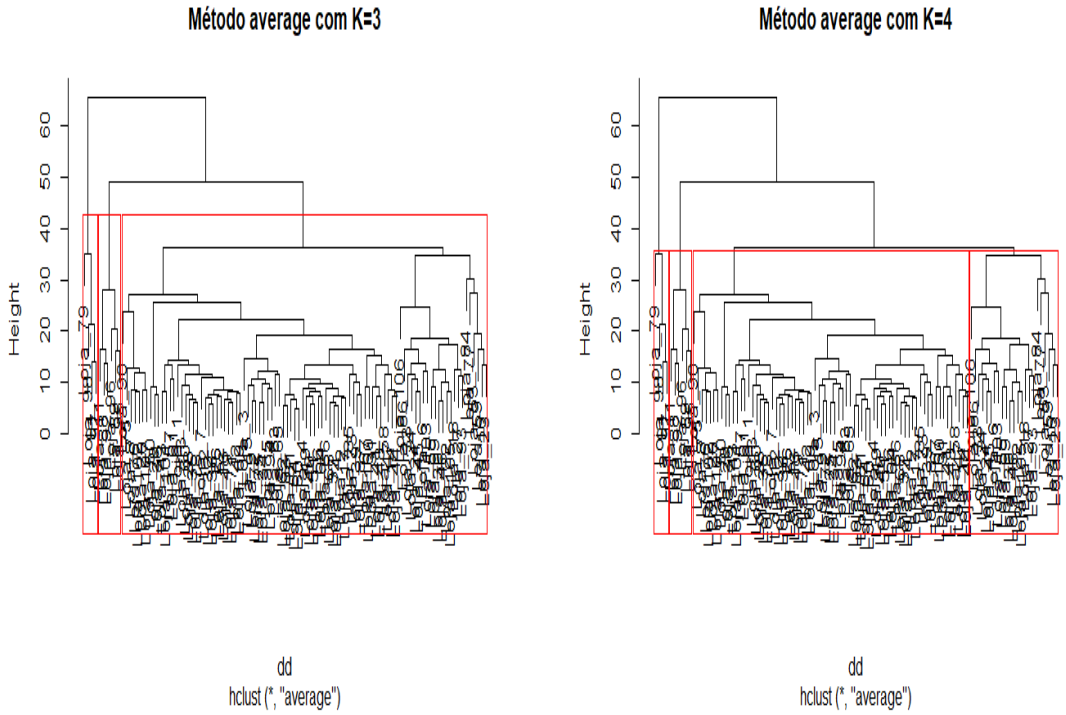

Figura 4.20: Representação da ligação average.

Observa-se através dos dendrogramas, que ao aplicar o método average, o agrupamento formado, tanto para $K = 3$ como para $K = 4$, não será um bom método a utilizar, uma vez que não ocorre homogeneidade dentro de cada grupo formado. Salienta-se que um bom método de agrupamento forma grupos cuja semelhança entre elementos no mesmo grupo é elevada e a semelhança entre elementos de grupos diferentes é reduzida. Conclui-se assim que o algoritmo hierárquico aglomerativo não é o melhor a aplicar à base de dados.

De seguida, analisou-se o algoritmo PAM, utilizando os parâmetros $K = 3$ e $K = 4$, para tal aplicou-se a função *pam()* da livraria *cluster*.

Com a aplicação do parâmetro $K = 3$ ao algoritmo PAM, obtiveram-se as seguintes partições das observações e respetiva silhueta:

	Cluster 1	**Cluster 2**	**Cluster 3**
Nº de lojas	5	60	39

Tabela 4.2: Particionamento das 104 lojas, aplicando o algoritmo PAM com o parâmetro $K = 3$.

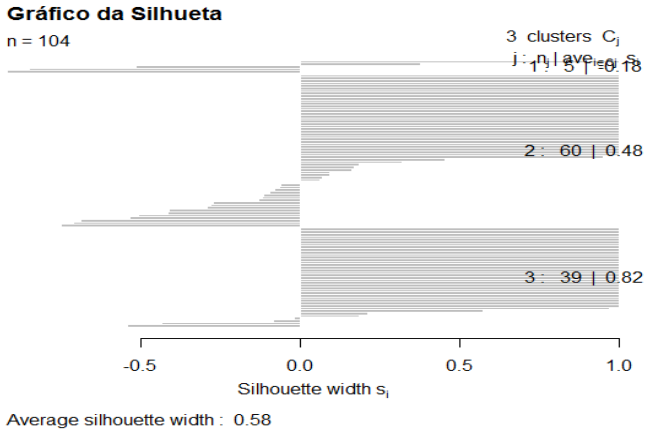

Figura 4.21: Representação da silhueta do método PAM com o parâmeto $K = 3$.

Pelo gráfico da silhueta, obtido através da função *plot(silhouette( ))*, que é uma técnica que avalia o particionamento das observações, verifica-se que existem 24 observações cujo valor da silhueta é negativo, ou seja, indica uma má alocação das lojas nos clusters. Contudo, o valor médio da silhueta é de 0.58, que indica que a estrutura é razoável.

Já na aplicação do parâmetro $K = 4$, ao algoritmo PAM, obtiveram-se as seguintes partições das observações e a respetiva silhueta:

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Nº de lojas	5	10	37	52

Tabela 4.3: Particionamento das 104 lojas, aplicando o algoritmo PAM com o parâmetro $K = 4$.

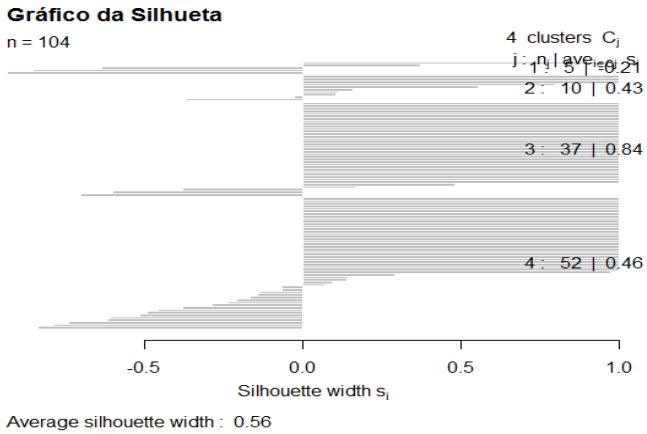

Figura 4.22: Representação da silhueta do método PAM com o parâmeto $K = 4$.

Pelo gráfico da silhueta verifica-se que existem 25 observações cujo valor da silhueta é negativo, ou seja, indica uma má alocação das lojas nos clusters. O valor médio da silhueta é de

0.56, que indica que a estrutura é razoável.

Conclui-se assim, que o algoritmo PAM não é um bom algoritmo a aplicar, uma vez que tanto para $K = 3$ como para $K = 4$, aproximadamente, 23% das lojas estão mal alocadas no cluster.

Por fim, aplicou-se o método K-médias para particionar as observações, de forma a agrupar as lojas semelhantes e as restantes ficarem em grupos distintos.

À semelhança dos métodos anteriores, aplicou-se $K = 3$ e $K = 4$ para obter o melhor particionamento das lojas. Para efetuar esta tarefa foi essencial a função *kmeans( )* da livraria *stats*. Com a aplicação do parâmetro $K = 3$ no algoritmo K-médias, obtiveram-se as seguintes partições e representações gráficas:

	Cluster 1	**Cluster 2**	**Cluster 3**
Nº de lojas	45	42	17

Tabela 4.4: Particionamento das 104 lojas, aplicando o algoritmo K-médias com o parâmetro $K = 3$.

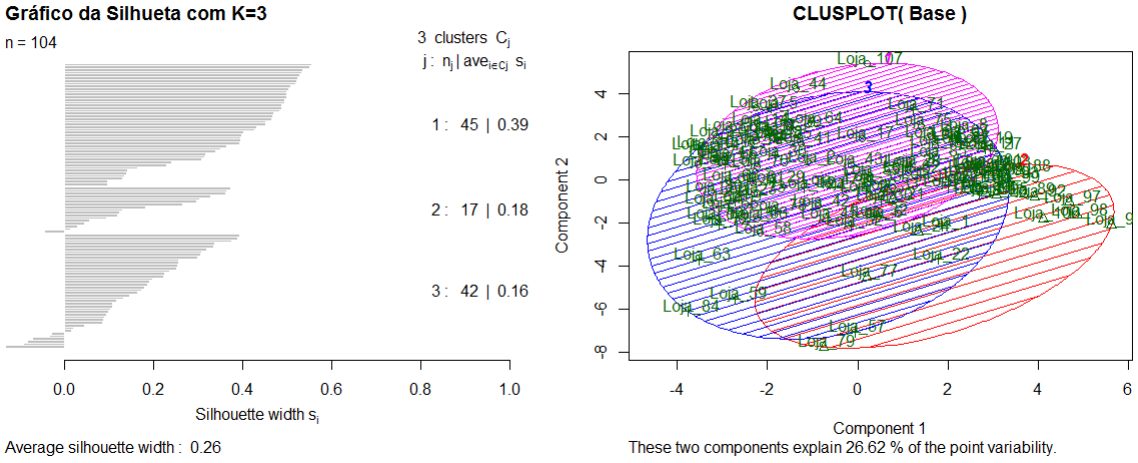

Figura 4.23: Representações da silhueta e do gráfico de clustering do método K-médias com o parâmeto $K = 3$.

Pelo gráfico da silhueta observa-se que existem 7 observações mal alocadas nos clusters. Mas o valor médio da silhueta é de 0.26 o que representa uma estrutura fraca. No gráfico dos clusters, obtido pela função *clusplot( )*, verifica-se que os três clusters parecem formar um bom particionamento das lojas.

Já na aplicação do parâmetro $K = 4$, ao algoritmo K-médias, obtiveram-se as seguintes partições das observações e os seguintes gráficos:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Nº de lojas	32	15	17	40

Tabela 4.5: Particionamento das 104 lojas, aplicando o algoritmo K-médias com o parâmetro $K = 4$.

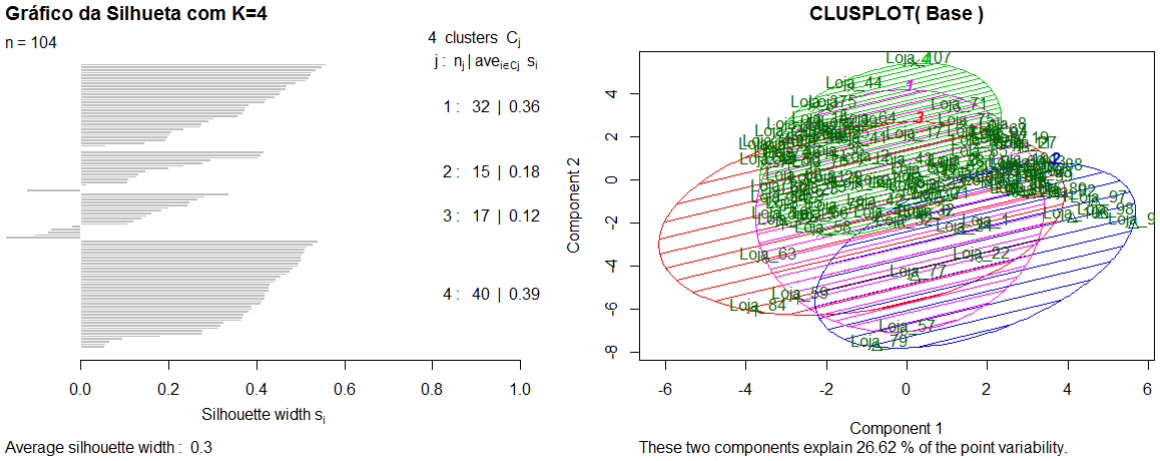

Figura 4.24: Representações da silhueta e do gráfico de clustering do método K-médias com o parâmeto $K = 4$.

Pelo gráfico da silhueta verifica-se que existem 7 observações cujo valor da silhueta é negativo. Mas o valor médio da silhueta é 0.3 o que representa uma estrutura fraca. Pelo *clusplot* observam-se clusters sobrepostos, o que não é uma representação ideal para os clusters.

Sendo assim, conclui-se que, o melhor algoritmo é o K-médias com o valor do parâmetro $K = 3$, uma vez que apenas 7 das 104 observações estão mal classificadas e pela representação do gráfico dos clusters constata-se um bom particionamento das lojas, pois nenhum dos clusters se sobrepõem. De forma a analisar o número de equipamentos em cada cluster, conclui-se que no cluster 1 contém 45 lojas, cujo número de equipamentos é reduzido. Já no cluster 2 existem 42 lojas, em que o número de equipamentos nelas é intermédio, enquanto que o cluster 3 é constituído por 17 lojas, cujo número de equipamentos é elevado comparativamente com as lojas nos restantes clusters. Com o motivo de analisar cada cluster, cruzou-se os dados com a área das lojas, obtendo-se os seguintes boxplots:

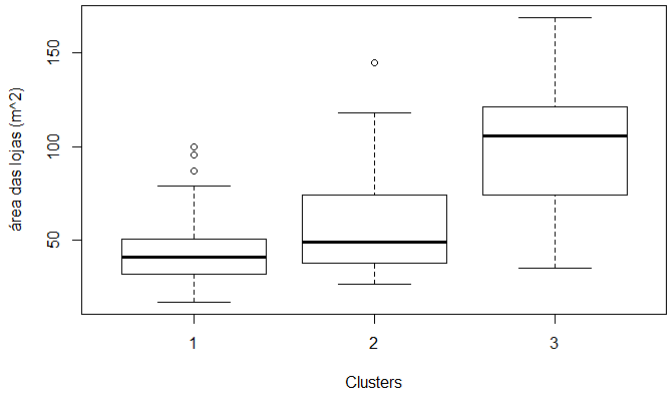

Figura 4.25: Representação dos boxplots referentes à área das lojas, em cada cluster.

Note-se que, o cluster 1 contém as lojas com a área menor e o cluster 3 contém, em média, as lojas em que a área é maior. Conclui-se, mais uma vez, que o método K-médias com o parâmetro $K = 3$ particionou bem as observações, apesar da estrutura ser fraca apenas 7 das observações estão mal alocadas e a partir do gráfico clusplot observa-se que não existe clusters sobrepostos.

De seguida, detetaram-se os outliers de cada cluster. Nas lojas retornadas como possíveis outliers, foi tido em conta o cluster a que pertence, o número de equipamentos, por metro quadrado (somou-se o número dos equipamentos, por loja, e dividiu-se pela sua área) e o valor dos equipamentos, por metro quadrado (somou-se o valor dos equipamentos, por loja, e dividiu-se pela sua área).

Na deteção de outliers são usadas técnicas não supervisionadas. Estas técnicas foram aplicadas nesta abordagem, a cada cluster, e cruzou-se com a informação da área de cada loja. É de relembrar que as observações foram agrupadas em três grupos: o do pouco equipamento, o do equipamento intermédio e o grupo das observações cujo número de equipamentos é maior.

Inicialmente, aplicaram-se técnicas baseadas em métodos estatísticos, mais concretamente, as não paramétricas. Das técnicas não paramétricas destaca-se o histograma, que ao nível do R é aplicado pelo comando *hist ()*, obteve-se a seguinte figura:

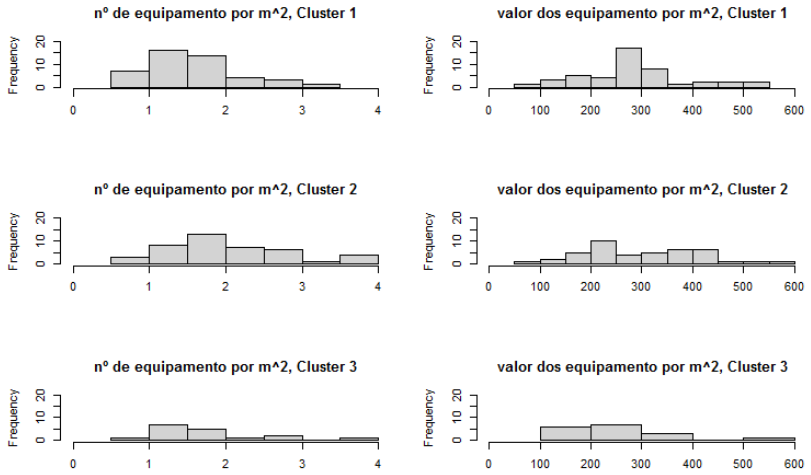

Figura 4.26: Representação do número de equipamentos e do valor dos equipamentos, por m^2, em cada cluster.

Observa-se que as lojas do cluster 1 contêm um número elevado de equipamentos, por metro quadrado, isto é, são lojas em que a área é reduzida mas têm um número elevado de equipamentos. Estas lojas são possíveis outliers. No cluster 1, observa-se também uma loja cujo número de equipamentos, por metro quadrado, é próximo de 0, poderá ser uma loja de abertura, por consequente ainda não tem todos os equipamentos registados em *SAP*. Já no cluster 3 verifica-se lojas cujo número de equipamentos e respetivo valor, por metro quadrado, é elevado.

De seguida, aplicaram-se técnicas de deteção de outliers baseadas em clustering, mais concretamente o algoritmo FindCBLOF. Este algoritmo particiona primeiro o conjunto de dados em clusters através do algoritmo K-médias, já aplicado anteriormente e posteriormente, calcula o valor do CBLOF para cada observação. As cinco lojas cujo valor do CBLOF, determinado pelo algoritmo FindCBLOF, é maior são: "Loja_6", "Loja_81", "Loja_53", "Loja_10" e "Loja_11".

Por fim, aplicaram-se as técnicas de deteção de outliers baseadas na proximidade, sendo estas classificadas em distância e densidade. Estas técnicas foram aplicadas nos três clusters. A técnica mais aplicada, baseada na distância, é K vizinhos mais próximos, KNN, que determina o score de outlier de uma observação com base nas distâncias dos K vizinhos mais próximos.

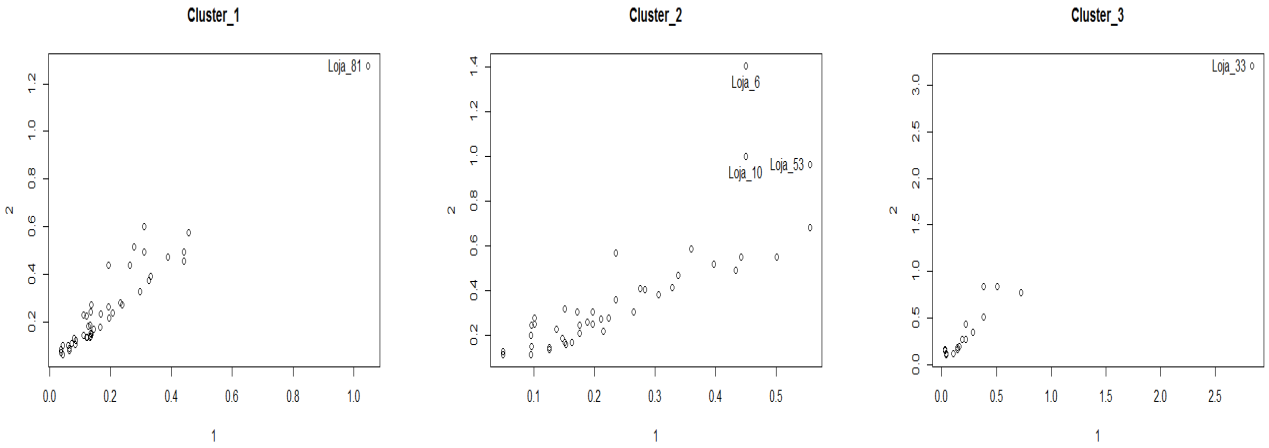

Figura 4.27: Aplicação do método KNN em cada cluster.

Para efetuar esta tarefa foi essencial a função *kNNdist( )* da livraria *dbscan*. Recorrendo ao método KNN conclui-se que no cluster 1 é evidente que a "Loja_81" é um outlier. Já no cluster 2 não é tão evidente a deteção de outliers, destacando-se apenas a "Loja_6", a "Loja_10" e a "Loja_53", pelo facto de se encontrarem mais afastadas. Enquanto que, no cluster 3, à semelhança do que ocorreu no cluster 1, a "Loja_33" destaca-se das restantes lojas. Para se obterem os gráficos 4.27Aplicação do método KNN em cada cluster. figure.caption.73, utilizou-se o parâmetro $K = 2$.

De seguida, aplicaram-se aos dados técnicas de deteção de outliers baseadas na densidade: LOF e LOCI.

A primeira técnica aplicada foi o LOF. Para determinar o valor de LOF, em *R*, recorreu-se à livraria *Rlof*. Está técnica foi aplicada a cada cluster. O valor de LOF foi determinado para cada observação, utilizando diferentes parâmetros de K (3, 6, 9, 12 e 15) pelo facto do cluster 3 ser constituído apenas por 17 observações. Obtiveram-se os seguintes gráficos, que representam o valor de LOF de cada observação, em cada cluster:

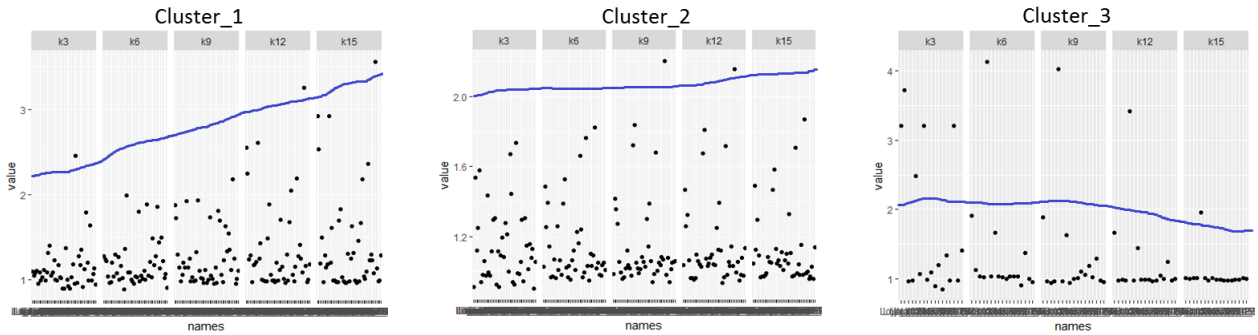

Figura 4.28: Aplicação do método LOF, em cada cluster.

No cluster 1 destaca-se a "Loja_66" e a "Loja_81". Já no cluster 2 foi mais complicado determinar os outliers, pois só no $K = 9$ e no $K = 12$ é que se destaca a "Loja_71". Para os restantes valores de K não existem outliers. Por fim, no cluster 3, para o parâmetro $K = 3$, as lojas consideradas foram: "Loja_106", "Loja_77", "Loja_96", "Loja_103" e "Loja_33". No cluster 3, à medida que se aumenta o 'MinPts', o valor de LOF tende a estabilizar, com a exceção da "Loja_33".

Por fim, foi aplicada a técnica LOCI. Esta técnica, ao contrário de LOF, não é necessário que seja definido um valor de corte, pois o próprio algoritmo determina quais as observações que são consideradas outliers. Na aplicação do LOCI foi necessário instalar a livraria *SMLoutliers*, para se obter as lojas outliers.

Na aplicação do LOCI no cluster 1, obteve-se a "Loja_74" e a "Loja_81". No cluster 2 obteve-se as seguintes lojas como possíveis outliers: "Loja_6", "Loja_10", "Loja_11" e "Loja_53". Já no cluster 3, obteve-se a "Loja_33", a "Loja_77", a "Loja_96" e a "Loja_103". Note-se que, foi considerado o parâmetro $alpha = 0.5$ devido a devolver um número considerável de lojas como possíveis outliers.

Na tabela E.1Lojas alarmísticas destacadas em cada técnica .table.caption.89, em anexo ETabelas da abordagem 2Anexo.a.E, apresenta-se de uma forma resumida as lojas destacadas em cada técnica de deteção de outliers, por cluster.

Conclui-se que, as lojas destacadas, nos clusters, são: "Loja_81", "Loja_6", "Loja_10", "Loja_11", "Loja_53", "Loja_96", "Loja_77", "Loja_103" e "Loja_33". Salienta-se ainda que, nesta abordagem, cruzou-se os dados com a variável área.

As lojas destacadas na tabela E.2Lojas alarmísticas obtidas na abordagem 2. table.caption.90, em anexo ETabelas da abordagem 2table.caption.89, são lojas que têm demasiados equipamentos para a sua dimensão. Por exemplo, a "Loja_33" é a loja que tem mais equipamentos e mais valor, mas como sofreu uma remodelação há pouco tempo, estas conclusões poderão dever-se ao facto de ainda não terem sido feitas as devidas movimentações. A "Loja_53" destaca-se por ser uma abertura e por já ter um número considerável de equipamentos para a sua área. A "Loja_77", a "Loja_96" e a "Loja_103" não são lojas alarmísticas, umas vez que, comparativamente com as restantes lojas da tabela E.2Lojas alarmísticas obtidas na abordagem 2. table.caption.90 os valores dos equipamentos destas lojas são reduzidos.

4.1.3 Abordagem 3

O objetivo do modelo é determinar as lojas cujo número de equipamentos, por metro quadrado, é reduzido ou elevado. Para tal, aplicou-se uma outra abordagem muito semelhante à

anterior. Nesta abordagem utilizaram-se as bases de dados, Base_1 e Data_1, que contêm 104 observações (lojas pertencentes à insígnia Y) e 40 variáveis (tipo de equipamentos cujo valor médio é superior a 30 €). O que diferencia da abordagem anterior é o facto de não se aplicar o clustering e apenas se analisar as lojas em geral.

O gráfico 4.29Representação da dispersão dos dados estandardizados. figure.caption.75 permite avaliar que, à medida que o valor dos equipamentos em cada uma das lojas, por metro quadrado, aumenta, no geral, o número de equipamentos, por metro quadrado, também aumenta.

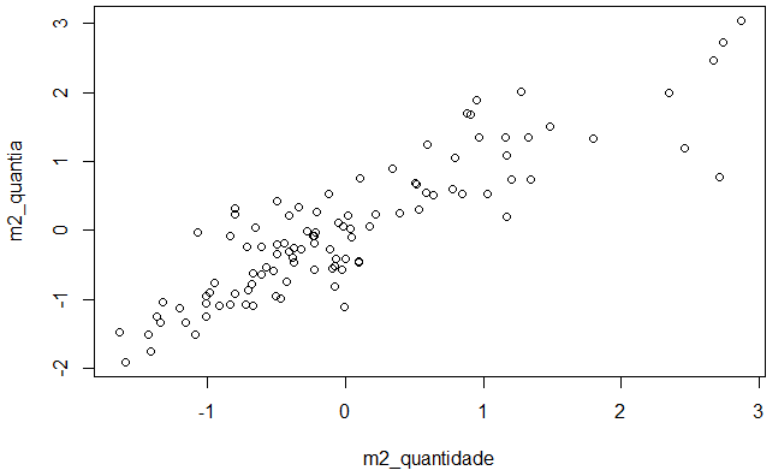

Figura 4.29: Representação da dispersão dos dados estandardizados.

Na deteção das lojas outliers são usadas técnicas não supervisionadas. Nesta abordagem, apenas aplicaram-se técnicas baseadas em métodos estatísticos e baseadas na proximidade.

Inicialmente, foram aplicadas técnicas baseadas em métodos estatísticos. Das técnicas não paramétricas destaca-se o boxplot, que a nível do R é aplicado pelo comando *boxplot ()*, obtendo-se as seguintes figuras:

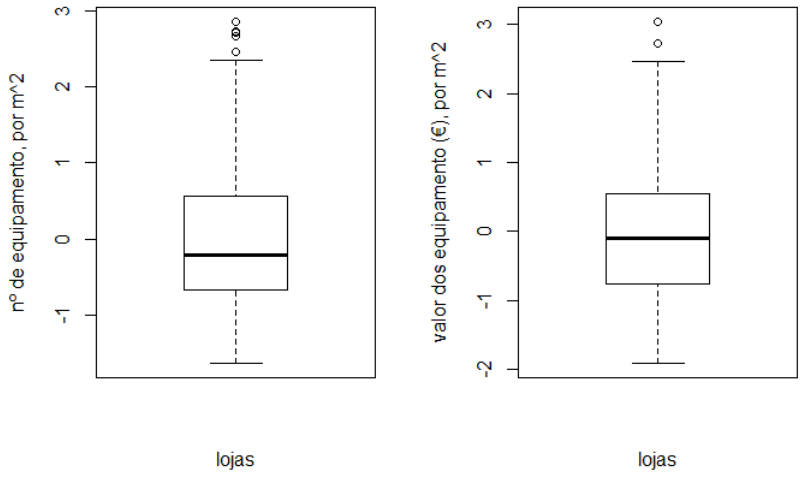

Figura 4.30: Representações do número de equipamentos e do valor dos equipamentos, por m^2.

Observando os boxplots, constata-se que existem lojas outliers, que contêm um número/valor elevado de equipamentos, por metro quadrado. No entanto, também existem lojas cujo número/valor de equipamentos, por metro quadrado, é reduzido. Estas lojas podem ter sido lojas de abertura e ainda não terem todos os equipamentos registados em *SAP*.

Por fim, foram aplicadas as técnicas de deteção de outliers baseadas na proximidade. A técnica baseada na distância mais aplicada é K vizinhos mais próximos, KNN.

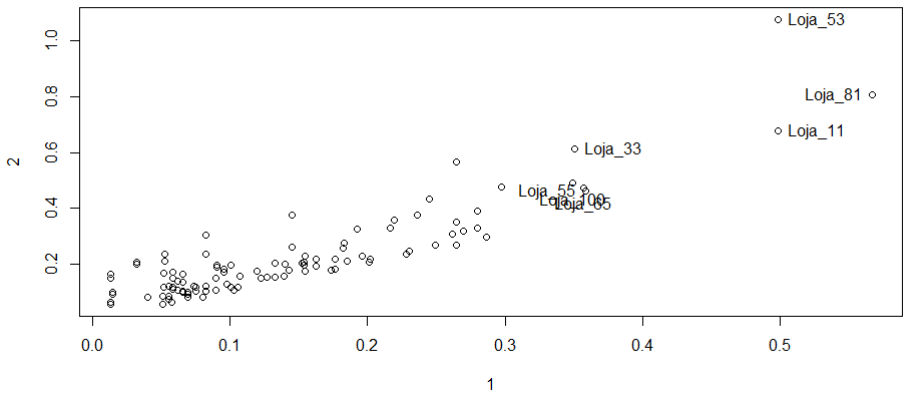

Figura 4.31: Representação do gráfico da aplicação da técnica KNN.

Pelo método KNN, determinou-se as seguintes lojas, como possíveis outliers: "Loja_53", "Loja_81", "Loja_11", "Loja_33", "Loja_55", "Loja_100" e "Loja_65". As lojas referidas destacam-se por terem uma distância aos seus K-vizinhos mais próximos maior do que as restantes lojas. Para a obtenção do gráfico 4.31Representação do gráfico da aplicação da técnica KNN.

figure.caption.77, utilizou-se o parâmetro $K = 2$, tendo sido testado, para o método KNN, diferentes K's obtendo-se sempre as mesmas conclusões.

Posteriormente, foram aplicadas aos dados, técnicas de deteção de outliers, baseadas na densidade: LOF e LOCI.

A técnica LOF foi a primeira a ser aplicada. O valor de LOF é maior num outlier de que numa observação 'normal'. Nesta técnica, aplicaram-se diferentes K's, (3, 6, 9, 12, 15, 20, 25, 30, 35 e 40). Salienta-se ainda que, nesta abordagem, são analisadas 104 observações, em conjunto. Representou-se o valor de LOF de cada observação e para cada K, num gráfico, determinado a partir do número e do valor dos equipamentos, por metro quadrado.

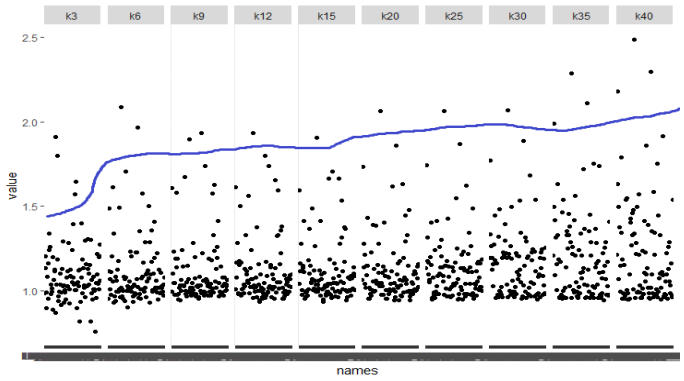

Figura 4.32: Aplicação do método LOF à base de dados em estudo.

Registou-se que, para $K = 3$ se obteve as seguintes lojas: "Loja_22", "Loja_25", "Loja_55" e "Loja_53". Para $K = 6$ obteve-se as lojas "Loja_25" e "Loja_53", já para $K = 9$ o resultado foram as lojas "Loja_53" e "Loja_33". A partir do $K = 12$ até ao $K = 30$ apenas tem-se a "Loja_33". Por fim, de destacar ainda, que quando $K = 35$ e $K = 40$, se obtiveram as lojas "Loja_33", "Loja_6" e "Loja_10". Desta forma, constata-se que as possíveis lojas outliers, retornadas pela técnica LOF, são: "Loja_22", "Loja_25", "Loja_55", "Loja_53", "Loja_33", "Loja_6" e "Loja_10".

Por fim, foi aplicada a técnica LOCI. Para esta técnica, não foi necessário definir um valor de corte, uma vez que o próprio algoritmo determina quais as observações que são consideradas outliers.

A loja que o valor LOCI apenas retornou como sendo um possível outliers foi a "Loja_53", com o parâmetro $alpha = 0.5$. Para o parâmetro alpha=0.8, resultaram, como possíveis outlires, as lojas "Loja_53", "Loja_11" e "Loja_81".

Na tabela F.1Lojas alarmísticas destacadas em cada técnica .table.caption.91, em anexo FTabelas da abordagem 3Anexo.a.F, apresenta-se de uma forma resumida as lojas destacadas em cada técnica de deteção de outliers. Analisando a tabela, destacam-se as lojas: "Loja_11",

"Loja_33", "Loja_53", "Loja_55" e "Loja_81". Salienta-se ainda que, nesta abordagem, cruzou-se os dados com a variável área.

As lojas destacadas na tabela F.2Lojas alarmísticas obtidas na abordagem 3. table.caption.92, em anexo FTabelas da abordagem 3table.caption.91, são lojas que, na abordagem 2, já tinham sido retornadas como possíveis outliers, com a exceção da "Loja_55"que, pelos valores obtidos, quer pelo número de equipamentos, por m^2, quer pelo valor, por m^2, não será uma loja relevante para a equipa de Inventariação.

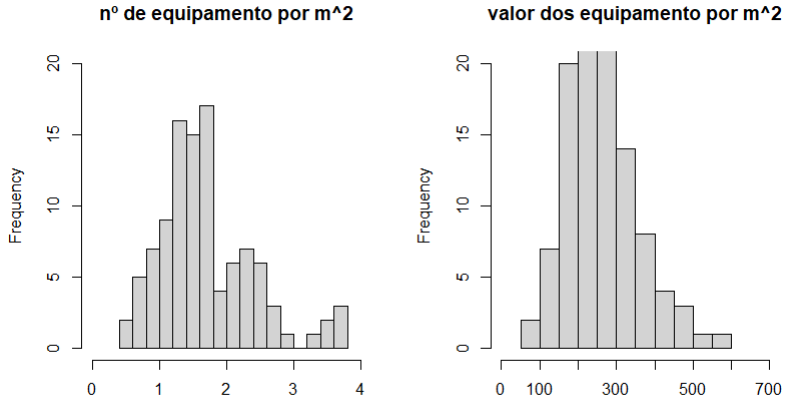

Figura 4.33: Representações do número de equipamentos e do valor dos equipamentos, por m^2.

Pelos gráficos da figura 5.1Representação resumida das lojas outliers obtidas nas três abordagens. figure.caption.80 conclui-se que as lojas com mais de 3 equipamentos, por m^2, são: "Loja_6", "Loja_10", "Loja_11", "Loja_33", "Loja_53" e "Loja_81". As lojas com mais de 500 €, por m^2, são: "Loja_6" e "Loja_33".

Assim sendo, pelos algoritmos aplicados, que as lojas alarmísticas, nesta abordagem, são: "Loja_6", "Loja_11", "Loja_33", "Loja_53" e "Loja_81".

Capítulo 5

Conclusão

De forma resumida as lojas obtidas em cada abordagem são:

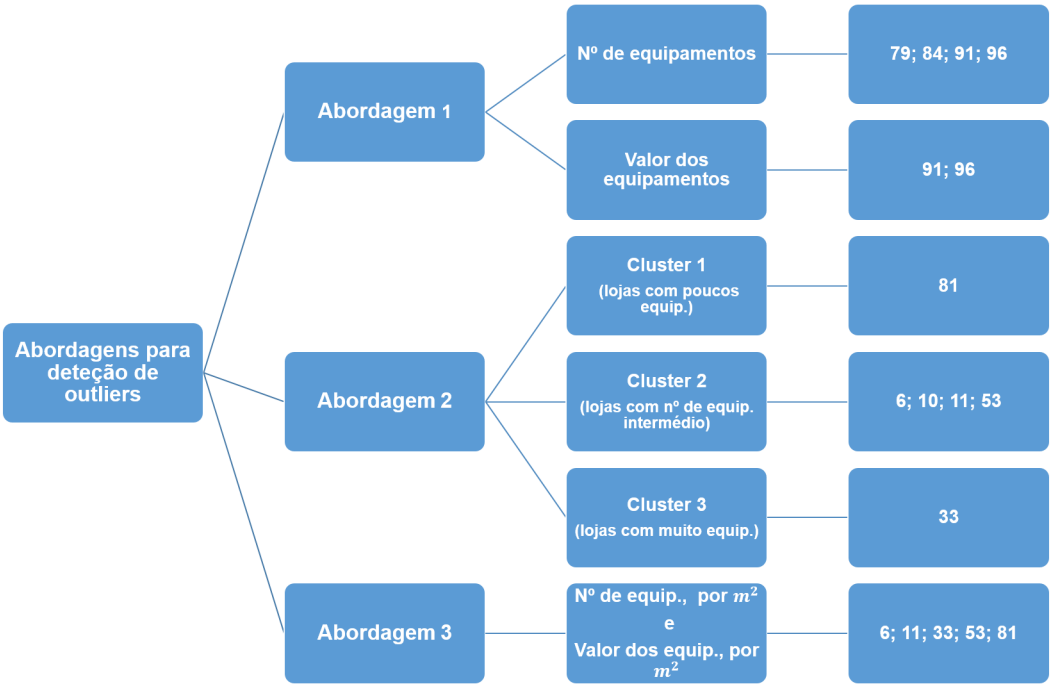

Figura 5.1: Representação resumida das lojas outliers obtidas nas três abordagens.

Perante os resultados, conclui-se que as lojas outliers que a abordagem 1 retornou, quer ao nível do número de equipamentos, quer ao nível do valor dos equipamentos, não estão em conformidade com as restantes abordagens. Já nas abordagens 2 e 3, podem-se destacar como possíveis lojas outliers, em comum, as seguintes lojas: "Loja_6", "Loja_11", "Loja_33", "Loja_53" e "Loja_81".

A equipa de "Gestão de Ativos Fixos" deverá ter em conta as seguintes lojas, uma vez que

o número de equipamentos e o valor dos equipamentos, por m^2, é elevado comparativamente às restantes lojas:

Lojas outliers	**Área**	**Nº de equipamentos por m^2**	**Valor dos equipamentos por m^2**	**Tipo de intervenção**
"Loja_6"	32 m^2	3.63	520.69 €	Remodelação em 2017
"Loja_11"	33.52 m^2	3.58	413.73 €	Remodelação em 2017
"Loja_33"	35 m^2	3.71	552.13 €	Remodelação em 2017
"Loja_53"	31.90 m^2	3.61	332.86 €	Abertura
"Loja_81"	21.20 m^2	3.35	451.28 €	Possível remodelação em 2017

Tabela 5.1: Top-5 das lojas alarmísticas.

Note-se que a "Loja_6", a "Loja_11" e a "Loja_33" são lojas que sofreram remodelação este ano e a equipa da Movimentação, que tem a função de dar indicação para abater ou transferir os equipamentos para outra loja, poderá ainda não ter feito as devidas movimentações dos equipamentos, sendo uma possível justificação para o facto das lojas terem muitos equipamentos. Já a "Loja_53" e a "Loja_81" são lojas que terão de ser devidamente analisadas.

Nas três abordagens, de entre os algoritmos aplicados às bases de dados, as técnicas KNN, LOF e o LOCI determinaram, quase sempre, as mesmas lojas. Já o algoritmo FindCBLOF destacava lojas como sendo possíveis outliers, mas eram lojas ditas 'normais' (não outliers).

No estágio, foi aplicada a abordagem 2, uma vez que está em consenso com as lojas alarmísticas, retornadas pela abordagem 3 e é também a abordagem mais completa ao nível de técnicas aplicadas. Na abordagem 2 aplicou-se o clustering, com o intuito de agrupar as lojas conforme o número de equipamentos e o seu valor. Na aplicação da técnica do clustering foram abordados métodos hierárquicos e não hierárquicos. De seguida, avaliou-se cada um dos métodos através do coeficiente de correlação cofonética e da largura média de silhueta. O método que se destacou, por particionar bem as observações, foi o K-médias, com o parâmetro $K = 3$. Seguidamente, aplicou-se a cada cluster técnicas de deteção de outliers não supervisionadas, para retornar as lojas mais atípicas. As técnicas aplicadas foram: FindCBLOF, KNN, LOF e LOCI. Nesta abordagem, as lojas obtidas em cada técnica, por cluster, são lojas alarmísticas.

A abordagem 2 foi inserida no software *R*, em que se utilizou o *R Markdown* com a finalidade de gerar um relatório de forma a facilitar as equipas da "Gestão de Ativos Fixos" a analisar as possíveis lojas outliers destacadas. Para tal, foi necessário instalar a livraria *rmarkdown*, que gera um relatório em formato *PDF*, *HTML* ou *Word*.

Salienta-se que o objetivo deste projeto foi analisar todas as insígnias pertencentes à SONAE, sendo que nesta tese apenas se referiu a insígnia *Y*, por ser das mais completas ao nível

da tipologia dos equipamentos. As abordagens aplicadas à insígnia Y foram também aplicadas às restantes insígnias da SONAE.

A SONAE, mais concretamente a equipa da "Gestão de Ativos Fixos", criou uma área partilhada com informação das lojas que o modelo alarmístico destacou, para partilha entre equipas, análise dos respetivos ativos e acompanhamento de eventuais ações. O modelo será executado anualmente, uma vez que se verificou que, num curto período de tempo, os bens das lojas não sofriam grandes alterações.

Futuras extensões deste trabalho, devem-se ao facto de incorporar outras abordagens, ou seja, em vez de ser aplicado o cluster à tipologia dos equipamentos, poder-se-ia ter agrupado as lojas pelo tipo de intervenção (abertura, remodelação e sem intervenção), pelos anos de existência, entre outros fatores. Em função da indisponibilidade de algumas informações, como a faturação ou a idade das lojas, poderiam ser consideradas variáveis que influenciar os modelos criados e por consequente alterariam as lojas destacadas como sendo lojas alarmísticas. Outra possível extensão seria aplicar a cada abordagem mais técnicas de deteção de outliers e comparar os resultados obtidos com as técnicas já estudadas.

Bibliografia

[1] Aggarwal C. (2016). ***Outlier analysis.*** 2ª edição, Springer, New York. Doi: 10.1007/978-3-319-47578-3

[2] Almeida R., Dias A. & Carvalho F. (2009). ***O Novo Sistema de Normalização Contabilística – SNC Explicado, ATF – Edições Técnicas***.

[3] Angiulli F. & Pizzuti C. (2002). ***Fast outlier detection in high dimensional spaces.*** In: Elomaa T., Mannila H. & Toivonen H. (eds.) PKDD 2002. LNCS (LNAI), Springer Berlin / Heidelberg, 2431º Volume, pp. 43–78. DOI: 10.1007/3-540-45681-3_2

[4] Barkan O. & Averbuch A. . ***Robust Subspace Mixture Models for Anomaly Detection in High Dimensions.*** IEEE Transactions on Journal NAME, Manuscript ID.

[5] Bhat A. (2014). ***K-Medoids clustering using partitioning around medoids for performing face recognition.*** International Journal of Soft Computing, Mathematics and Control (IJSCMC),3ª edição, 3º Volume.

[6] Breunig M., Kriegel H., Ng R. & Sander J. (2000). ***LOF: Identifying Density-Based Local Outliers.*** In: ACM SIGMOD Record, ACM, pp. 93-104. DOI: 0.1145/335191.335388

[7] Chandola V., Banerjee A. & Kumar V. (2009). ***Anomaly Detection: A survey.*** ACM Computing Surveys (CSUR), pp. 1-38. DOI: 10.1145/1541880.1541882

[8] Chandola V., Boriah S. & Kumar V. (2008). ***Similarity Measures for Categorical Data: A comparative evaluation.*** Depeartment of Computer Science and Engineering, University of Minnesota. In:SDM, SIAM, Philadelphia, pp. 243-253. DOI: 10.1137/1.9781611972788.22

[9] Duda O., Hart E. & Stork G. (2001). ***Pattern classification.*** 2ª edição, John Wiley & Sons, New York, NY.

[10] Eduardo H., Estevam Jr. & Nelson E. (2003). ***A Nearest-Neighbor Method as a data preparation tool for a clustering genetic algorithm.*** In Proceedings of the 18th Brazilian Symposium on Databases / ACM SIGMOD Disk (SBBD 2003) Manaus: Editora da Universidade Federal do Amazonas, pp. 319-327.

[11] Ehmke J. (2012). ***Integration of information and optimization models for routing in city logistics.*** Springer Science & Business Media, 177º Volume de International Series in Operations Research & Management Science, pp. 50.

[12] Ester M., Kriegel H., Sander J. & Xu X.(1996). ***A Density-Based Algorithm for discovering clusters in large spatial databases with noise.*** In: Proceedings of second international conference on knowledge discovery and data mining, Portland, Oregon, pp. 324-331. DOI: 10.1.1.121.9220

[13] Exemplo do algoritmo LOF. Visto a Junho 8, 2017 de `http://www.cse.ust.hk/~leichen/courses/comp5331/lectures/LOF_Example.pdf`.

[14] Fayyad U., Piatetsky-Shapiro G. & Smyth P.(1996). ***From Data Mining to knowledge discovery in databases.*** 3ª edição, AI Magazine, 17º Volume, pp. 37-54.

[15] Goldschmidt R. & Passos E. (2015). ***Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações.*** 2ª edição, Elsevier, Rio de Janeiro.

[16] Han J., Pei J. & Kamber M. (2011). ***Data Mining: Concepts and techniques.*** - The Morgan Kaufmann Series in Data Management Systems. 3ª edição, Elsevier, USA, pp. 543-580.

[17] Hautamäki V., Kärkkäinen I. & Fränti P. (2004). ***Outlier detection using k-nearest neighbour graph.*** University of Joensuu, Department of Computer Science Joensuu, Finland. DOI: 10.1109/ICPR.2004.1334558

[18] Hawkins D. (1980). ***Identification of outliers.*** - Monographs on Statistics and Applied Probability. Springer Science & Business Media, South Africa. DOI: 978-94-015-3994-4

[19] He Z., Xu X. & Deng S. (2003). ***Discovering cluster-based local outliers.*** Pattern Recognition Letters 24, pp. 1641-1650. DOI: 10.1016/S0167-8655(03)00003-5

[20] Knorr M. & Raymond Ng (1998). ***Algorithms for mining distance-based outliers in large datasets.*** In:Proc. Int. Conf. on Very Large Databases (VLDB 1998), pp. 392-403.

[21] Kopka J., Reves M. & Giertl J. (2010). ***Anomaly detection techniques for adaptive anomaly driven traffic engineering.*** Dept. of Computers and Informatics, FEEI TU of Košice, Slovak Republic, (SCYR 2010), 10th Scientific Conference of Youn Researchers.

[22] Kohonen T. (1998). ***The self-organizing map.*** Helsinki University of Technology, Neural Networks Research Centre, Elsevier. DOI: 10.1016/S0925-2312(98)00030-7

[23] Navega S. (2002). ***Princípios Essenciais do Data Mining.*** Publicado nos Anais do Infoimagem, Cenadem.

[24] Palestino C.(2011). ***BI2-Business Intelligence: modelagem e qualidade.*** Elsevier, Brasil.

[25] Papadimitriou S., Kitagawa1 H., Gibbons P. & Faloutsos C. (2003). ***LOCI: Fast outlier detection using the local correlation integral.*** IEEE, 19th International Conference on Data Engineering (ICDE'03), Bangalore, India. DOI: 10.1109/ICDE.2003.1260802

[26] Piegorsch W. (2015). ***Statistical Data Analytics: foundations for Data Mining, informatics, and knowledge discovery, dolutions.*** John Wiley & Sons, Journal of Statistical Software, 69º Volume, Chichester.

[27] Primak V. (2008). ***Decisões com BI (Business Intelligence).*** Ciência Moderna, Fabio Vinicius Primak, pp. 33-34.

[28] Ramasmawy R., Rastogi R. & Kyuseok S. (2000). ***Efficient algorithms for mining outliers from large data sets.*** Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, TX. DOI: 10.1145/342009.335437

[29] Rao A. & Srinivas V. (2008). ***Regionalization of Watersheds: An approach based on cluster analysis.*** Springer Science & Business Media, 28º Volume, pp. 28-29.

[30] Rcoreteam R: A Language and Environment for Statistical Computing. Vienna, Austria, 2013. Disponível em: `http://www.R-project.org/`.

[31] Rousseeuw J.(1987). ***Silhouettes: A graphical and to the interpretation and validation of cluster analysis.*** Elsevier Science Publisher B. V., Journal of Computational and Applied Mathematics, 20º Volume, North-Holland, pp. 53-65. DOI: 10.1016/0377-0427(87)90125-7

[32] Rousseuw J., Ruts I. & Tukey W. (1999). ***The bagplot: A bivariate boxplot.*** The American Statistician, pp. 382–387. DOI: 10.1080/00031305.1999.10474494

[33] Smiti A. & Elouedi Z. (2012). ***DBSCAN-GM: An improved clustering method based on gaussian means and DBSCAN techniques.*** In: Proc. of the IEEE 16th International Conference on Intelligent Engineering Systems (INES), pp. 13–15. DOI: 10.1109/INES.2012.6249802

[34] Sokal R. & Rohlf J. (1962). ***The comparison of dendrograms by objective methods.*** International Association for Plant Taxonomy (IAPT). Taxon, Berlin, Volume 11, number 2, pag. 30-40. DOI: 10.2307/1217208

[35] Tang J., Chen Z., Fu A. & Cheung D. (2002). ***Enhancing effectiveness of outlier detections for low density patterns.*** In Chen, M.-S., Yu, P., Liu, B. (eds.)PAKDD 2002. LNCS, 2336º Volume, Springer Berlin / Heidelberg Advances, pp. 535–548.

[36] Torgo L. (2010) ***Data Mining with R: Learning with Case Studies.*** Chapman & Hall.

[37] Xu R. & Wunsch D. (2008). ***Clustering.*** John Wiley & Sons, 10º Volume de IEEE Press Series on Computational Intelligence, pp. 32.

Anexos

Anexo A

Exemplo da aplicação do algoritmo LOF

Para um melhor entendimento da deteção de outliers baseado na densidade, inclui-se o seguinte exemplo [13]. Considera-se os 4 pontos que se seguem: $a(0,0)$, $b(0,1)$, $c(1,1)$ e $d(3,0)$ (figura A.1Representação dos pontos. figure.caption.84), para cada ponto calcula-se o valor de LOF e determina-se o top-1 outlier, usando $K = 2$ e a distância Euclidiana.

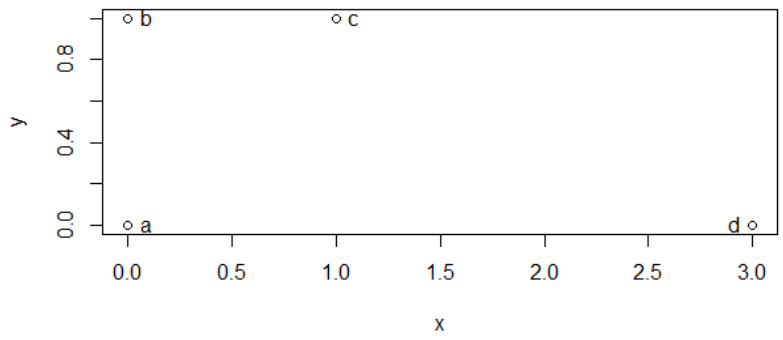

Figura A.1: Representação dos pontos.

1. A distância Euclidiana dos pontos é: $dist(a,b) = 1$; $dist(a,c) = \sqrt{2} \simeq 1.41$; $dist(a,d) = \sqrt{9} = 3$; $dist(b,c) = 1$; $dist(b,d) = \sqrt{10} \simeq 3.16$ e $dist(c,d) = \sqrt{5} \simeq 2.24$.

2. A K-distância de cada observação, com $K = 2$, é: $2 - \text{distância}(a) = dist(a,c) \simeq 1.41$ (c é o segundo vizinho mais próximo); $2 - \text{distância}(b) = dist(b,a) = 1$ (a ou c é o segundo vizinho mais próximo); $2 - \text{distância}(c) = dist(c,a) \simeq 1.41$ (a é o segundo vizinho mais próximo) e $2 - \text{distância}(d) = dist(d,a) = 3$ (a é o segundo vizinho mais próximo).

3. A vizinhança K-distância de cada observação, para $K = 2$, é: $N_{2-\text{distância}(a)}(a) = \{b, c\}$; $N_{2-\text{distância}(b)}(b) = \{a, c\}$; $N_{2-\text{distância}(c)}(c) = \{a, c\}$ e$N_{2-\text{distância}(d)}(d) = \{a, c\}$.

4. A densidade de alcance local de cada observação, é:

$$dal_2(a) = \frac{\mid N_2(a) \mid}{max\{2-\text{distância}(b), dist(a,b)\} + max\{2-\text{distância}(c), dist(a,c)\}} = \frac{2}{max\{1,1\} + max\{1.41, 1.41\}} = \frac{2}{1+1.41} = 0.83;$$

$$dal_2(b) = \frac{\mid N_2(b) \mid}{max\{2-\text{distância}(a), dist(b,a)\} + max\{2-\text{distância}(c), dist(b,c)\}} = \frac{2}{max\{1.41,1\} + max\{1.41, 1\}} = \frac{2}{1.41+1.41} = 0.71;$$

$$dal_2(c) = \frac{\mid N_2(c) \mid}{max\{2-\text{distância}(a), dist(c,a)\} + max\{2-\text{distância}(b), dist(c,b)\}} = \frac{2}{max\{1.41,1.41\} + max\{1, 1\}} = \frac{2}{1.41+1} = 0.83;$$

$$dal_2(d) = \frac{\mid N_2(d) \mid}{max\{2-\text{distância}(a), dist(d,a)\} + max\{2-\text{distância}(c), dist(d,c)\}} = \frac{2}{max\{1.41,3\} + max\{1.41, 2.24\}} = \frac{2}{3+2.24} = 0.38.$$

5. O valor de LOF de cada observação, é:

$$LOF_2(a) = \frac{\frac{dal_2(b)}{dal_2(a)} + \frac{dal_2(c)}{dal_2(a)}}{\mid N_2(a) \mid} = \frac{\frac{0.71}{0.83} + \frac{0.83}{0.83}}{2} = 0.93;$$

$$LOF_2(b) = \frac{\frac{dal_2(a)}{dal_2(b)} + \frac{dal_2(c)}{dal_2(b)}}{\mid N_2(b) \mid} = \frac{\frac{0.83}{0.71} + \frac{0.83}{0.71}}{2} = 1.17;$$

$$LOF_2(c) = \frac{\frac{dal_2(a)}{dal_2(c)} + \frac{dal_2(b)}{dal_2(c)}}{\mid N_2(c) \mid} = \frac{\frac{0.83}{0.83} + \frac{0.71}{0.83}}{2} = 0.93;$$

$$LOF_2(d) = \frac{\frac{dal_2(a)}{dal_2(d)} + \frac{dal_2(c)}{dal_2(d)}}{\mid N_2(d) \mid} = \frac{\frac{0.83}{0.38} + \frac{0.83}{0.38}}{2} = 2.18.$$

6. Ordena-se de forma decrescente todos os valores de LOF, determinados anteriormente: $LOF_2(d) = 2.18$; $LOF_2(b) = 1.17$; $LOF_2(a) = 0.93$ e $LOF_2(c) = 0.93$.

7. O top-1 outlier é o ponto d, pois é o que tem maior valor de LOF em comparação com os restantes pontos.

Anexo B

Descrição das variáveis do conjunto de dados

Variável	Descrição
Nº inventário	Código da etiqueta.
Empr	Nome da insígnia da base de dados.
Cen_Custo	Nome da loja.
Descritivo Centro de custo	Se é loja ou se é um armazém/SEDE.
Desc_família	Tipologia do equipamento.
Desc_sub-família	Descrição do equipamento.
Localização	Onde se encontra o equipamento.
Desc_local.	Descrição do local onde se encontra o equipamento.
Tipo de equipamento	O tipo de equipamento.
Dta invent	Data em que foi feita a auditoria.
Estado do bem	Estado em que se encontra o equipamento.
Imobilizado	Corresponde ao ativo da empresa, tem uma ordem sequencial por empresa.
Sbnº	É uma componente ou adição ao ativo, por defeito é zero (Ex: cadeira – imobilizado 1, subnº. 0).
Dt Aq.Ori.	Data em que a loja adquiriu o equipamento.
Ini. Dpr. no	Data em que o equipamento iniciou a depreciação.
Qtd	Quantidade do bem.
Família	Agregador das sub-famílias e indica qual o grupo de equipamentos a que cada sub-família pertence. Ex: 02 – Equipamento de Frio
Sub-Família	Indica o equipamento a um nível mais desagregado. Ex: 0201 – Expositores de frio
UBM	Significa a que unidade de medida pertence a quantidade.
Det. ctas	É a 'conta' do imobilizado para agregação das fichas de imobilizado, por conta.
Cófigo DGCI	É o código fiscal atribuído pela AT e publicados nas tabelas anexas ao DR 25/2009 (classificação fiscal dos bens de imobilizado).
AA-Benefícios	É um código que quando atribuído aos bens, significa qual o benefício em que esses bens foram enquadrados Ex: DD – REFAI – Regime Fiscal de Apoio ao Investimento.
Fornecedor	Fornecedor do bem.
Val.aquis.atual	Valor da aquisição atual do bem.
Moeda \ Moeda1	Moeda - euro (€).
Val.cont.fi.exer.	Valor contabilístico fixo do exercício.
Observações	Comentários importantes sobre o equipamento.
Fabricante do equipamento	Código de quem fabricou o bem.
Data movimentação Permitida	Data quando foi permitida a venda \ abate do bem.

Tabela B.1: Descrição das variáveis .

Anexo C

Algoritmo FindCBLOF

```
area$cluster_base<-numeric(nrow(Base))
area$cluster_base<-kmeans(Base,3)$cluster # Aplicar o clustering
 var1 <- nrow(area[which(area[,5]==(order(table(area$cluster_base))[3])),]); #nº de elementos
 no cluter maior
var2 <- (nrow(area[which(area[,5]==(order(table(area$cluster_base))[3])),])
/nrow(area[which(area[,5]==(order(table(area$cluster_base))[2])),]));
    D <- nrow(area); #numero de observações
    i <- 3; LC <- (order(table(area$cluster_base))[i]);
    alpha=0.75
    beta=3
  while (var2 < beta & var1 < (alpha*D)){
    i = i - 1;
    var1 <- var1 + nrow(area[which(area[,5]==(order(table(area$cluster_base))[i])),]);
    var2 <- (nrow(area[which(area[,5]==(order(table(area$cluster_base))[i])),])
     /nrow(area[which(area[,5]==(order(table(area$cluster_base))[i-1])),]));
    LC <- c(LC, order(table(area$cluster_base))[i]) }
     SC <- c();
    for(i in (1:3)){
      if(sum(i == LC) == 0){ SC <- c(SC, i) }}
     m<-nrow(area)
     area$CBLOF<-numeric(m)
    for (i in 1:nrow(area)){
        if(sum(1 == LC) == 0){
          if (area[i,5]==1){
              area$CBLOF[i]<-nrow(area[which(area[,5]==1),])
               *min(dist(c(area[i,1], min(dist(area[which(area[,5]==3),1])))),
               dist(c(area[i,1],min(dist(area[which(area[,5]==2),1]))))} }
      if(sum(2 == LC) == 0){
        if (area[i,5]==2){
```

```
        area$CBLOF[i]<-nrow(area[which(area[,5]==2),])
        *min(dist(c(area[i,1], min(dist(area[which(area[,5]==3),1])))),
        dist(c(area[i,1], min(dist(area[which(area[,5]==1),1]))))}}
    if(sum(3 == LC) == 0){
      if (area[i,5]==3){
        area$CBLOF[i]<-nrow(area[which(area[,5]==3),])
        *min(dist(c(area[i,1], min(dist(area[which(area[,5]==1),1])))),
        dist(c(area[i,1],min(dist(area[which(area[,5]==2),1]))))}}
    if(sum(1 == SC) == 0){
      if (area[i,5]==1){
        area$CBLOF[i]<-nrow(area[which(area[,5]==1),])*dist(c(area[i,1],
         min(dist(area[which(area[,5]==1),1])))} }
    if(sum(2 == SC) == 0){
      if (area[i,5]==2){
        area$CBLOF[i]<-nrow(area[which(area[,5]==2),])*dist(c(area[i,1],
         min(dist(area[which(area[,5]==2),1])))}}
    if(sum(3 == SC) == 0){
      if (area[i,5]==3){
        area$CBLOF[i]<-nrow(area[which(area[,5]==3),])*dist(c(area[i,1],
        min(dist(area[which(area[,5]==3),1])))} } }
row.names(area[order(area$CBLOF, decreasing=T)[1:5],])
```

Anexo D

Tabelas da abordagem 1

Técnicas	Lojas outliers
FindCBLOF	29; 62; 71; 4; 15
KNN	84; 96; 91; 79
LOF	84; 96; 91
LOCI	79; 84; 89; 91; 96; 97; 98

Tabela D.1: Lojas alarmísticas obtidas na base de dados Base .

Técnicas	Lojas outliers
KNN	96; 91
LOF	50; 40; 91; 96; 88
LOCI	79; 84; 91; 96; 105; 106

Tabela D.2: Lojas alarmísticas obtidas na base de dados Data .

Lojas outliers	Nº de equipamentos	Valor dos equipamentos	Tipo de intervenção
"Loja_84"	178	26317.38 €	Remodelação em 2016
"Loja_91"	246	43051.74 €	Remodelação em 2016
"Loja_96"	276	42616.89 €	Remodelação em 2017
"Loja_79"	244	31458.88 €	Remodelação em 2017

Tabela D.3: Lojas alarmísticas obtidas na abordagem 1.

Anexo E

Tabelas da abordagem 2

Cluster	FindCBLOF	KNN	LOF	LOCI
Lojas do Cluster_1 (lojas com pouco equipamento)	81	81	66; 81	74; 81
Lojas do Cluster_2 (lojas com equipamento intermédio)	6; 53; 10; 11	6; 10; 53	71	6; 10; 11; 53
Lojas do Cluster_3 (lojas com muito equipamento)	-	33	106; 77; 96 ; 103; 33	33; 77; 96; 103

Tabela E.1: Lojas alarmísticas destacadas em cada técnica .

Lojas outliers	Área	Nº de equipamentos por m^2	Valor dos equipamentos por m^2	Tipo de intervenção
"Loja_6"	32 m^2	3.78	559.04 €	Remodelação em 2017
"Loja_10"	31.90 m^2	3.61	519.09 €	Abertura
"Loja_11"	33.52 m^2	3.57	413.73 €	Remodelação em 2017
"Loja_33"	35 m^2	3.94	595.06 €	Remodelação em 2017
"Loja_53"	31.90 m^2	3.64	355.68 €	Abertura
"Loja_77"	63.50 m^2	2.50	328.46 €	Remodelação em 2017
"Loja_81"	21.20 m^2	3.49	494.56 €	Possível remodelação em 2017
"Loja_96"	87 m^2	2.16	308.37 €	Remodelação em 2017
"Loja_103"	42 m^2	2.64	363.17€	Abertura

Tabela E.2: Lojas alarmísticas obtidas na abordagem 2.

Anexo F

Tabelas da abordagem 3

Técnicas	Lojas outliers
KNN	53; 81; 11; 33; 55; 100; 65
LOF	22; 25; 55; 53; 33; 6; 10
LOCI	53; 11; 81

Tabela F.1: Lojas alarmísticas destacadas em cada técnica .

Lojas outliers	Área	Nº de equipamentos por m^2	Valor dos equipamentos por m^2	Tipo de intervenção
"Loja_11"	33.52 m^2	3.58	413.73 €	Remodelação em 2017
"Loja_33"	35 m^2	3.94	594.06 €	Remodelação em 2017
"Loja_53"	31.90 m^2	3.64	355.68 €	Abertura
"Loja_55"	40.40 m^2	2.65	304.48 €	Remodelação em 2017
"Loja_81"	21.20 m^2	3.49	494.56 €	Possível remodelação em 2017

Tabela F.2: Lojas alarmísticas obtidas na abordagem 3.